FINAL REPORT

PARALLEL READOUT OF OPTICAL DISKS

Submitted to:

U.S. ARMY RESEARCH OFFICE
P.O. Box 12211
Research Triangle Park, NC 27709

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 31, 1992	Final 15 June 1989 - 14 June 1992

4. TITLE AND SUBTITLE
Parallel Readout of Optical Disks

5. FUNDING NUMBERS
DAAL03-89-K-0114

6. AUTHOR(S)
Demetri Psaltis

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
California Institute of Technology
Pasadena CA 91125

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 26676.15-PH

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Optical disks can be utilized in a variety of systems that take advantage of the inherent parallel accesability to the stored data. We experimentally demonstrate the use of parallel access to data stored on optical disks in digital optical computing, several types of neural networks including optical and optoelectronic neural networks, image classifiers, and image correlators.

14. SUBJECT TERMS

15. NUMBER OF PAGES
80

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

PARALLEL READOUT OF OPTICAL DISKS

FINAL REPORT

PROFESSOR DEMETRI PSALTIS

31 AUGUST 1992

U.S. ARMY RESEARCH OFFICE

CONTRACT/GRANT NUMBER DAAL03-89-K-0114

California Institute of Technology
Pasadena, California 91125

FOREWARD

This report is composed of reprints of papers published over the course of this program. The first paper, "Mass Storage for Digital Optical Computing," discusses the suitability from an architectural point of view of planar optical media and thick holographic optical media as mass memory in a digital optical computer. The capability of optical memory disks in particular are covered in the second paper, "Optical Memory Disks in Optical Information Processing." In this paper, characteristics such as contrast, diffraction efficiency, and phase uniformity are measured, and conclusions about the performance of optical disks in various architectures are given. For example, the paper covers parallel readout of data stored as either images or holograms stored on the disk. The use of optical disks in neural networks is introduced in the second paper and covered in depth in the third, "Optical-disk Based Artificial Neural Systems." The third paper describes experimental results from using an optical disk in two different neural network architectures. The first system uses the disk to store and implement the neural network connections in an optical character recognition system. In a second experiment a feed forward neural network reads connection patterns in parallel from an optical disk for implementation on an optoelectronic chip. Another neural network system that uses optical disks is described in the fifth paper, "Optical Implementation of Radial Basis Classifiers." In this system the disk stores reference radial basis functions which are read off in parallel by the neural network classifier. The fourth paper, "Image Correlators using Optical Memory Disks," describes an experimental demonstration of an image correlator that not only uses an optical disk to store a large library of images, but also as the spatial light modulator in the correlator.

Mass storage for digital optical computers

Demetri Psaltis, Alan Yamamura, and Hsin-Yu Li

California Institute of Technology
Department of Electrical Engineering
Pasadena, CA 91125

ABSTRACT

Optical memory and computing technologies have progressed significantly in the last few decades. In this paper, we review some of the current planar and 3-D optical memories and discuss how well they meet the requirements for mass memory in digital optical computers.

1. INTRODUCTION

A computer consists of nonlinear processing elements (*e.g.*, switches), interconnections, memory, and input/output peripherals. Most of the work in optical computing, including the digital approach, has concentrated on the switches, interconnections, and how these can be put together into architectures that perform useful tasks. Tanguay has pointed out the need to consider all the elements that comprise a computer, including the memory and I/O, in order to design overall systems that can have practical impact[1]. It is interesting to ponder why so little attention has been given recently to memories with parallel access capability, a subject of intense study in the sixties and early seventies[2]. Among all the uses of optics in computers, memory seemed to be the most likely to have an early impact, and in a way, this has proven to be correct with the advent of optical disks. However, largely due to material and device limitations, parallel optical memories never became practical. Perhaps, it is precisely this early focus on optical memories, which did not succeed, that discouraged continued research in this area, whereas the relatively new areas of optical switches and interconnections received most of the attention of researchers.

It is now time to reexamine parallel optical memories, in particular their role in optical computers. There are several reasons for this:

a. The optical device technology (optical switches, spatial light modulators, detectors, holograms, *etc.*) has progressed dramatically since the early seventies, and this has created new possibilities.

b. To a lesser extent, materials that can be used as the storage medium, have advanced. Photorefractives, organic materials, and magnetooptic media are still the prime candidates (as they were twenty-five years ago), but in some cases the understanding of these materials has increased considerably.

c. Optical computing has progressed significantly. As optical computers start to become practical, they will demand a parallel memory with possibly a huge storage capacity. A neural network is an example of a massively par-

allel computer architecture that maps well to optics and typically requires a very large mass memory with rapid access.

In this paper we will do three things. First we will discuss in general terms the requirements for a mass memory in a digital optical computer. We will then outline some of the optical memories that can be implemented with planar storage media, and finally we will discuss 3-D storage in volume holograms.

2. MASS STORAGE MEMORY

In almost all modern computers there are two types of memories: random access (RAM) and mass storage. In serial, von Neumann computers, the RAM is modified by a single central processing unit, and it stores the program that is currently being executed (or part of it), the data, and intermediate results as well as the operating system (or the necessary part of it). The mass memory (disk and tape) stores everything else. The capacity of the mass memory is much larger than the RAM, but information is retrieved and stored at much lower rates. RAM and mass memories are also part of parallel computers. The organization of the memory is a much more difficult task in a parallel computer, however. The RAM is organized either as "shared memory," which allows processors can use, or it is distributed throughout a "multiprocessor architecture," with each processor having its own memory. The mass memory in parallel architectures is standard disk or tape memory. Such systems are typically intended to operate without transfers to and from the mass memory during the execution of a program, since this would slow down the process. Therefore, the role of mass memory is usually not emphasized in the design of such architectures, even though in practice, mass memory is often a very serious bottleneck. A notable exception is database machines, which are specifically designed to search quickly through mass memory.

Digital optical computers are generally considered to be very fine-grain, massively parallel processors. The basic components of a digital optical computer are shown in the block diagram of Fig. 1. It consists of optical gates, interconnections, and mass memory. The optical gates are physically 2-D arrays of nonlinear optical switches. They can be arranged in a single or multiple planes. These gates are used to construct the processing elements (PEs) as well as the RAM for the system. The interconnections between these gates specify their functionality and the architectural design of the system. In some highly dedicated architectures (*e.g.*, cellular arrays for image processing) there may be little need for RAM memory, and the bulk of the available gates can be devoted to processors. However in most applications, the majority of the gates will have to be devoted to the memory function to store the program and intermediate results. The third component, the mass memory, is interfaced to the RAM so that it can load its programs and data as needed. In Fig. 1, a double line connecting two blocks indicates parallel transfer of information, whereas a single line represents a serial link. Notice that all links are parallel except for the transfer from the RAM to the mass memory. The low sensitivity of materials that might be used for fabricating the mass memory makes it very unlikely that a massively parallel (several thousand channels) parallel write memory can he made practi-

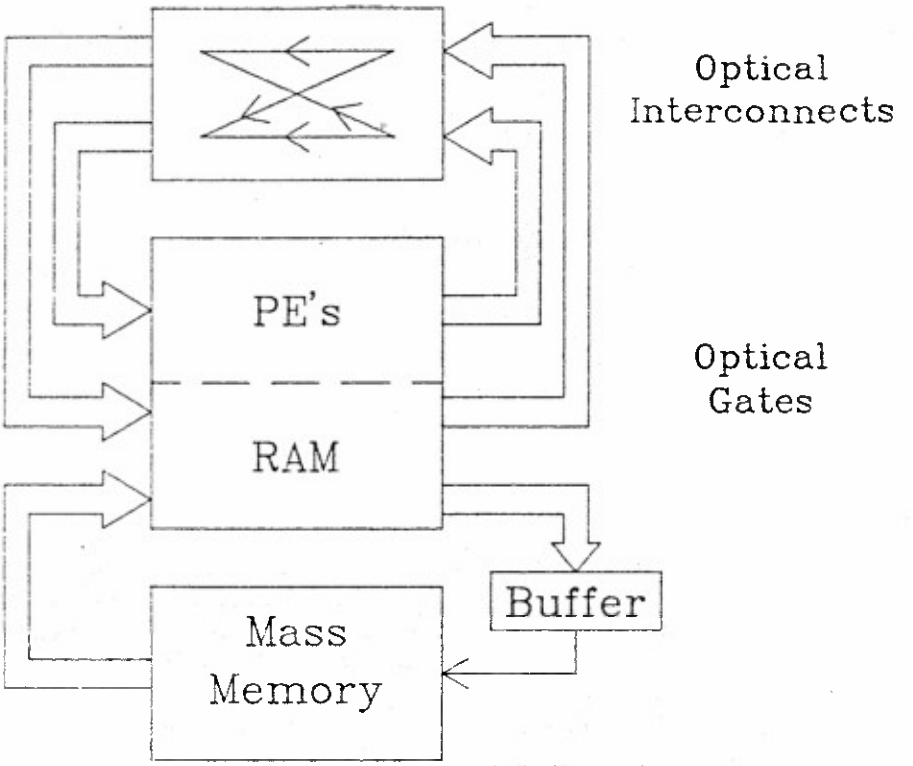


Fig. 1. Basic architecture of a digital optical computer.

cal (*i.e.*, fast enough). However, it is possible to have an optical mass memory that has parallel readout. Notice that in an electronic implementation this link would also be essentially serial. It is not clear how significant parallel access to mass memory is, it may represent a new target of opportunity for optical computing. The specifications for a parallel readout optical memory may be as follows:

- Large capacity (much larger than the capacity of the RAM).
- Parallel readout (10^3–10^6 channels).
- Fast access.
- Low probability of error $P_e \approx 10^{-5}$.

In the following two sections we will consider planar and 3-D media and comment on the prospects of the various memory systems.

3. 2-D PARALLEL READOUT OPTICAL MEMORIES

In this section we will discuss optical memories constructed with a planar medium. Such memories have been investigated intensively, for more than twenty-five years. There are a multitude of reasons why such memories have not yielded practical success. One of them is probably the fact that there has not

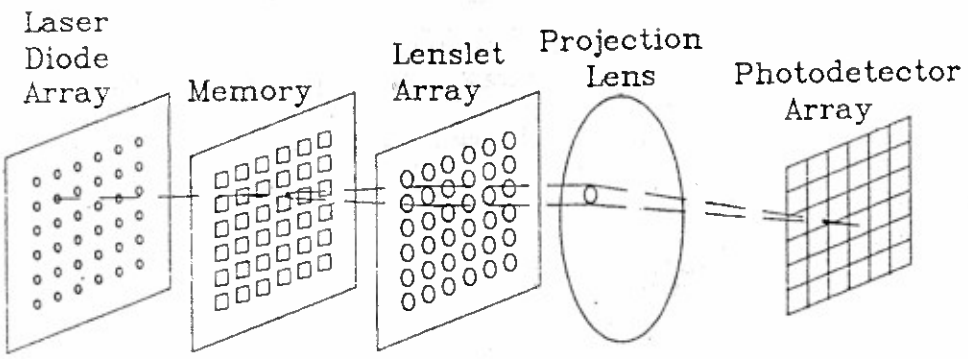


Fig. 2. Example of a 2-D optical memory system.

been a clear, urgent need for them. In other words, present computers do not really have the throughput to deal with the data rates that have been projected for tbese memories. Optical computers, if they come to maturation, should be able to make use of the capabilities of parallel optical memories. Another motivation for reexamining such memories is the emergence of optical disks that are used as serial memories. These same disks can also be used as parallel memories[3] thereby providing a mature technology for recording very accurate, computer controlled memories.

There is a wide variety of 2-D optical memory architectures. Reference 2 is an excellent review of early efforts in this area. Even though the details of the various architectures vary significantly, most share two basic characteristics: the data is organized in 2-D blocks or pages, and a scanning mechanism is used to address one of these blocks and transfer it to the output. We will describe one of these architectures as an example. This architecture is described in Reference 2, and it is reproduced here as Fig. 2. The memory is a 2-D transparency on which the blocks of data are recorded on a regular 2-D grid. Each of the blocks is addressed by a separate laser diode from a corresponding 2-D grid of laser diodes. The laser array serves as the scanner in this system. Each of the blocks is imaged to a common output plane with a pair of lenses arranged in the standard $4-f$ configuration. The first lens is part of a lenslet array, one lenslet per page. The second is a common large lens. The memory is addressed by turning on only one laser diode at a time. In this way only one of the blocks is transferred to the output. This memory has specifications that can probably be made compatible with the requirements set forth in the previous section, with one important exception: storage capacity. The total number of bits (number of pages $\times$ number of bits per page) that can be stored in such a system is limited primarily by the passive optics of the system (big lens and lenslet array). A reasonable estimate for a practical system is probably $10^8 = 10^4 \times 10^4$ bits. This would require a 100×100 lenslet array, each lenslet having a 100×100 pixels space-bandwidth product. This relatively small capacity is the biggest drawback of this type of memory. It might be argued that through careful engineering and clever design one might be able to design a system with the very rapid access of the system in Fig. 2 (limited only by the sensitivity of the

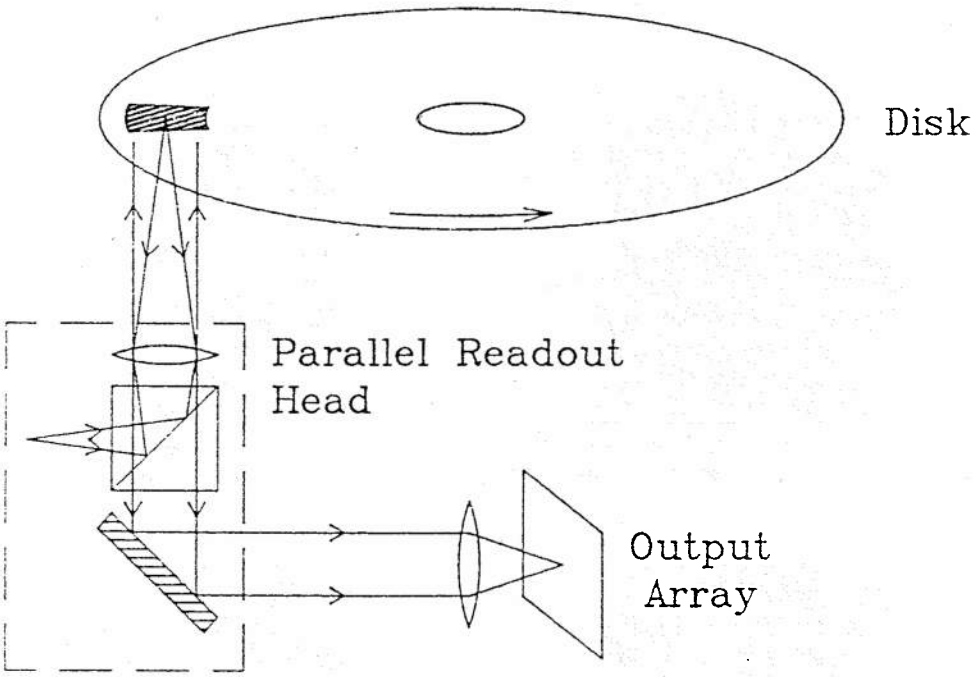


Fig. 3 Parallel readout version of an optical disk memory system.

output detector) but with larger capacity. For instance, since the lenses seem to be the major limitation in Fig. 2, a lensless system can be designed using holography to try to overcome this limitation. Unfortunately, analysis of the holographic version of this system[4] shows that it is not practical to increase the capacity with holography.

The speed of the memory in Fig. 2 is derived from the fact that each laser diode addresses a separate block or page. Thus the access time becomes equal to the time it takes to turn one of the lasers on. All these systems that rely on an optical scanning mechanism have such fast access time but limited capacity because the optical system/scanner needs to accommodate the *entire* space-bandwidth product of the memory. An alternative and complementary addressing mode is mechanical alignment. This is the method currently used in all serial mass memories including optical disks. A parallel readout version of this is shown in Fig. 3. Data is recorded on the disk, again on a 2-D array of 2-D blocks or pages. The blocks are centered on a polar grid around the disk. The disk rotation and the motion of the readout head in the radial direction align one of the blocks on the disk with the readout head, and the block is then imaged onto an output 2-D array. The space-bandwidth product of the optics is only equal to that of a single block (10^4–10^6 pixels) and hence can be easily constructed. The capacity of the memory is equal to the number of bits that can be stored on the disk (10^{10} bits or more). The disadvantage of this type of memory is relatively slow speed because of the need for mechanical motion ($10 - 100$ *msec* access time to any block).

Perhaps the biggest practical problem that needs to be worked out with a parallel access memory based on mechanical alignment is registration. The pixel size on the disk is approximately 1 μm and we need to register the image on the disk to the output detector array with that tolerance. The fact that we need to do this while the disk is spinning as fast as possible (to reduce the access time) makes the problem difficult. One possible solution to this problem is the use of Fourier transform holograms[3]. Instead of directly recording the block of data on the disk, we can instead record the 2-D Fourier transform of the data rather than an image of the data. Thus, the original block of data will be reconstructed at the output. The advantage of this method is that a small shift in the pattern recorded on the disk will not affect the reconstructed image because of the shift invariance property of the Fourier transform. This can simplify greatly the registration problem. The drawback of the holographic method is the large increase in the number of pixels on the disk that are needed to store the block.

Since the optical disks we use store only unipolar binary amplitude information, a space-bandwidth product (SBP) penalty must be paid to record holograms on the disks. With only real information written on the disk, we must record both the desired information and its complex conjugate, thus automatically doubling the SBP used. In addition, with only unipolar information stored on the disk, the reconstructions typically have large DC spots, easily doubling again the SBP to accomodate a spatial carrier that moves the desired part of the reconstruction away from DC. Finally, with only one bit of dynamic range per pixel, a significant amount of noise appears spread throughout the reconstruction. Depending on the desired number of pixels and signal-to-noise ratio (SNR) desired in the reconstruction, an additional factor of 4 to 256 or more in disk SBP could easily be required. Fig. 4 shows the desired object and experimental reconstruction from computer generated holograms recorded on the disk as we vary the number of "on" pixels in the reconstruction. As the number of "on" pixels in the reconstruction increases from an initial value of 32, the SNR decreases steadily. With 512 "on" pixels, the signal is almost completely lost in the noise. Over a million pixels on the disk were used to encode the 1024 pixels in the reconstruction, corresponding to an additional factor of 1024 in required SBP for holographic recording when compared with imaging.

4. 3-D STORAGE

We saw in the previous section that there is a basic trade-off between speed and storage capacity when a 2-D storage medium is used. This impasse can to some extent be broken when a 3-D storage medium is used[5]. The classic method for storing a number of pages in a 3-D hologram is shown in Fig. 5. A hologram of each of the pages is formed with a plane wave reference. The angle of the reference beam is unique to each page. Each page can be reconstructed by illuminating the recorded hologram with its own reference. Cross talk is eliminated because of the angular selectivity of a thick hologram. The storage capacity of such a module can be 10^9– 10^{10} bits (10^3 blocks with 10^6–10^7 pixels each). The access time to each block is determined by the time it takes to scan the reference beam from one angle to another (a few microseconds or less). Notice that

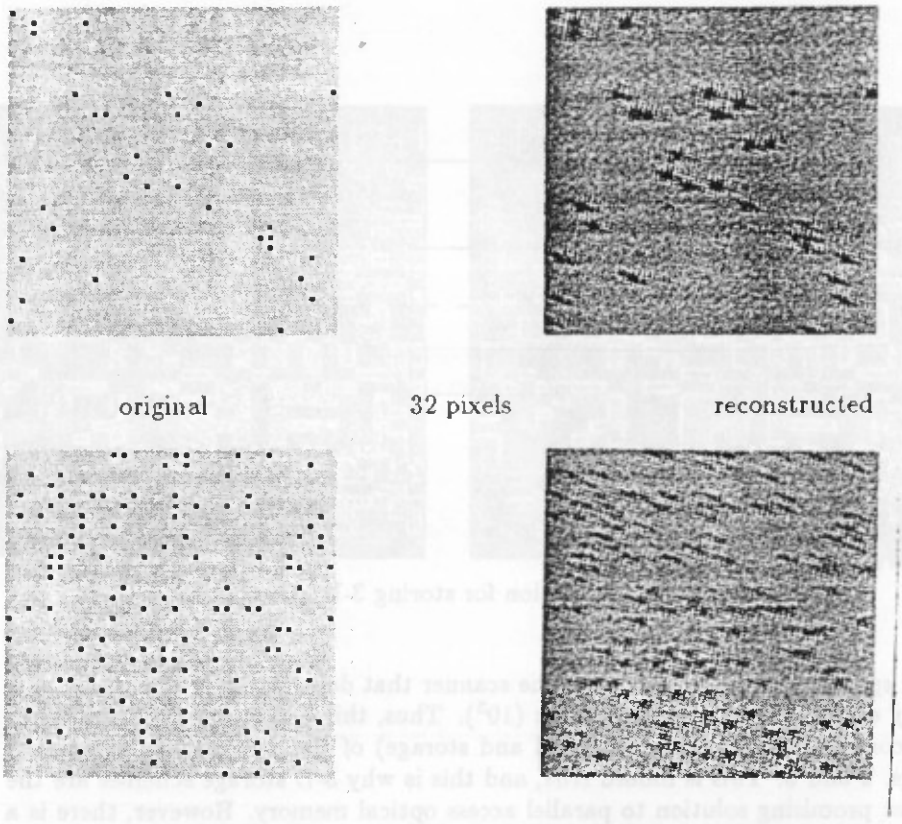


Fig. 4 Desired objects and experimental reconstructions of CGHs on disks with 32, 128, and 512 pixels on respectively.

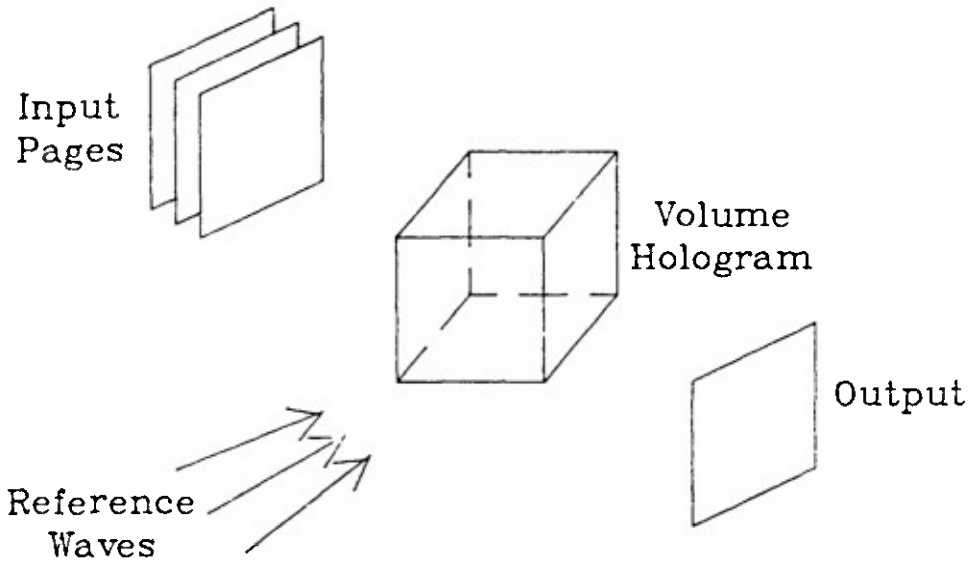
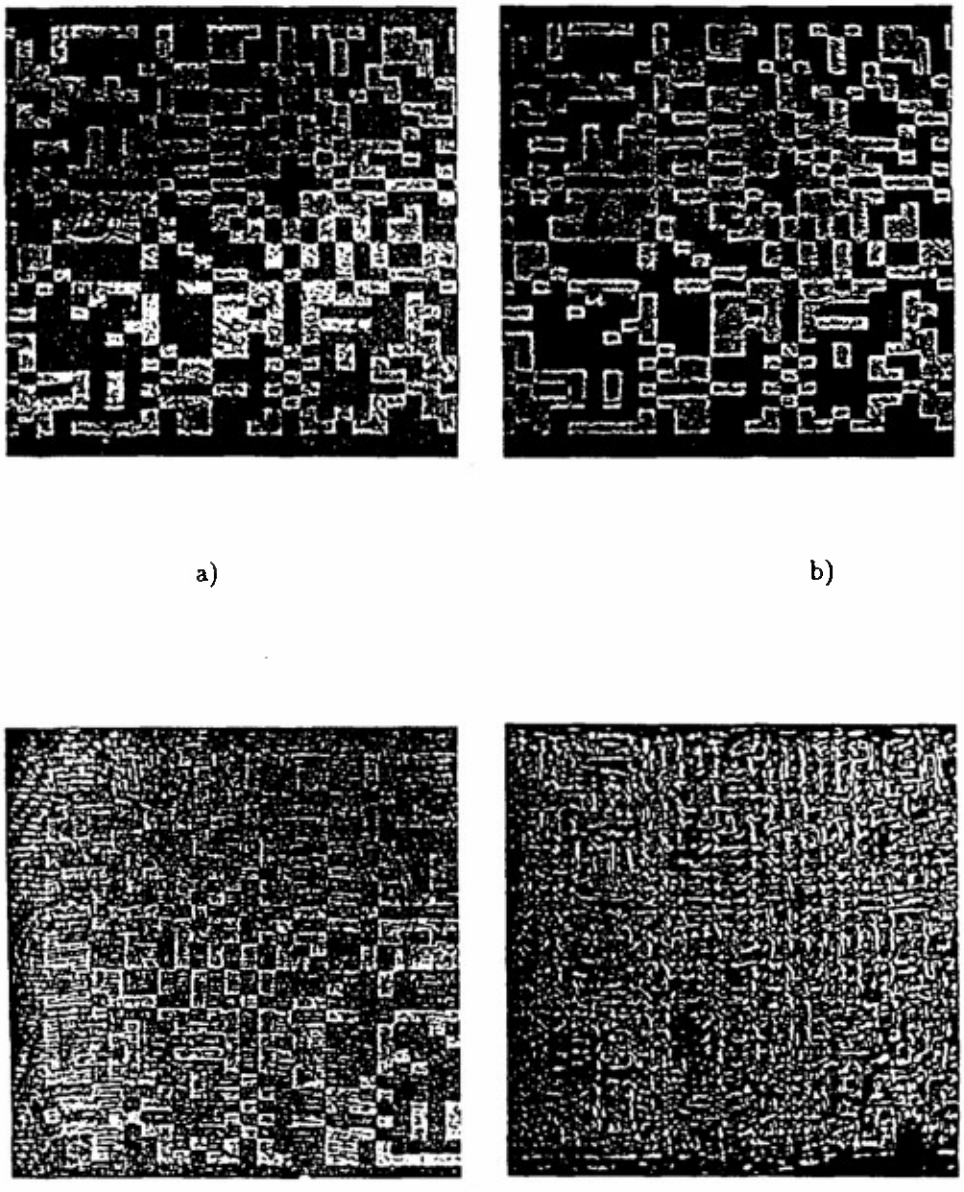


Fig. 5. Configuration for storing 3-D holograms.

the space-bandwidth product of the scanner that deflects the reference beam is only equal to the number of pages (10^3). Thus, this 3-D storage scheme seems to combine the advantages (speed and storage) of the 2-D storage schemes of Figs. 2 and 3. This is indeed true, and this is why 3-D storage schemes are the most promising solution to parallel access optical memory. However, there is a serious problem with 3-D storage which is a consequence of the fact that the pages must be recorded through a sequence of holographic exposures. A further consequence of this is a requirement for a dynamic recording medium. Problems that arise when multiple exposures are made on a simple crystal include reduction of the diffraction efficiency of each individual hologram[6], fanning, and nonlinear intermodulation terms.

Fig. 6a shows a binary pattern that is recorded as a single page onto a $LiNbO_3$ crystal, according to the system described in Fig. 5. After the pattern is written on the crystal, we can reconstruct it (read it out) with the reference beam, as shown in Fig. 6b. As new patterns (additional pages) are written onto the crystal, the original pattern decays in diffraction efficiency leading to the situation shown in Fig. 6c, where the diffracted light is comparable to noise due to fanning, scattered light, detector noise, *etc.* Because the old information stored on the crystal gets erased as new information is added, we try to achieve the maximal diffraction efficiency when writing the single images. However, if we try to write to saturation, the image quality deteriorates because of fanning and nonlinear intermodulation. Fig. 6d shows the result when we try to write the original pattern to saturation on the crystal. The highly distorted image is due to intermodulation effects. In practice, it is not advisable to write a single image to full saturation.

Despite such problems several experiments have been performed that store



a) b)

c) d)

Fig. 6 a) Original pattern, b) Readout with diffraction efficiency 0.01%, c) Readout with diffraction efficiency 0.0017%, d) Saturation.

up to 500 images[7,8] in single photorefractive crystals. One of the most challenging and important steps in this area is the development of materials and recording techniques that exhibit read-write and write-erase asymmetries which will allow us to better control the recording of information in three dimensions.

5. CONCLUSION

Table I summarizes our work by comparing the relative merits of nonmechanical scanning, mechanical scanning (optical disks), and 3-D storage (volume holograms) in terms of their speed, density, and accuracy. Optical disks are presently the most practical form of optical mass storage; they are technologically mature and provide high density and accuracy at the cost of relatively long access times. Nonmechanical scanning systems, like paged imaging and holographic memories, are plagued by low density and mediocre accuracy. Volume holography can provide high speed and density, but it is currently in the developmental stage. However, if continuing research in both materials and recording techniques achieves success, volume holography may represent the best technology in the future for mass storage in digital optical computers. Finally, the above techniques may be combined to create new hybrid memories. Figure 7, for example, shows a volume holographic optical disk that provides the high density of volume holograms with the large capacity and simplified scanning of the optical disk. (A Bragg cell provides optical scanning in the radial direction, complementing the mechanical scanning in the azimuthal direction.) Thus by utilizing the best aspects of several different technologies, hybrid memories could be developed that simultaneously provide high speed, density, capacity, and accuracy.

Table 1. Optical mass storage media comparison.

	Nonmechanical Scanning Systems	Mechanical Alignment (Disks)	3-D Storage
Speed	High	Low	High
Density	Low	High	High
Accuracy	Medium	High	Low
Practicality	Low	High	Medium

6. ACKNOWLEDGEMENTS

Thanks to Seiji Kobayashi for providing the computer generated holograms written on the optical disk. This work is supported by the AFOSR, ARO, and DARPA. Alan Yamamura is supported in part by a fellowship from the Fannie and John Hertz Foundation.

7. REFERENCES

1. A.R. Tanguay, Jr., "Material requirements for optical processing and computing devices," *Optical Engineering*, pp. 2-18, Vol. 24, No. 1, 1985.

2. J.P. Huignard, F. Micheron, and E. Spitz, "Optical systems and photo-

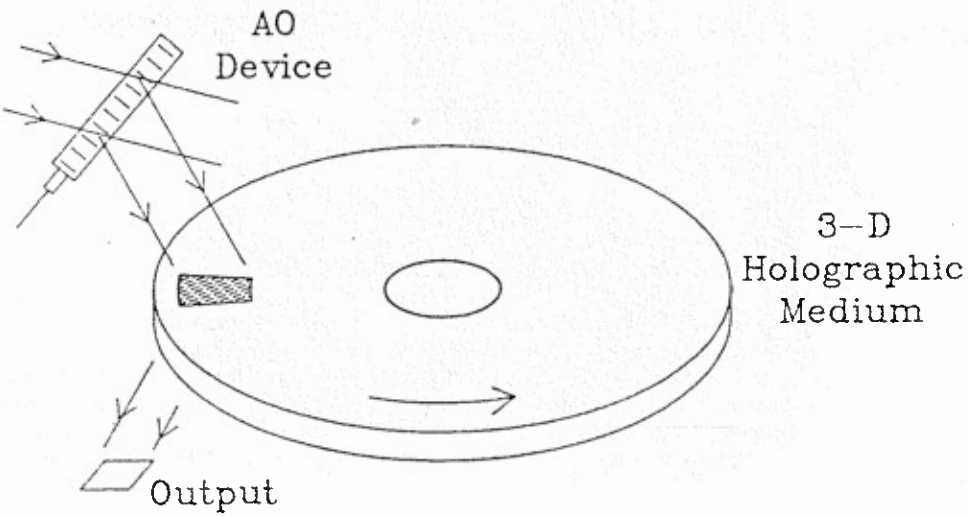


Fig. 7. Volume holographic optical disk.

sensitive materials for information storage," *Optical Properties of Solids*, B.O. Seraphin, Ed., Ch. 16, pp. 847–925, North Holland, Amsterdam, 1976.

3. D. Psaltis, M.A. Neifeld, A.A. Yamamura, and S. Kobayashi, "Optical memory disks in optical information processing," to appear in *Applied Optics*, May 1990 special issue.

4. F.M. Smits and L.E. Gallaher, "Design Considerations for a Semipermanent Optical Memory," *The Bell System Technical Journal*, pp. 1267–1278, Vol. XLVI, No. 6, 1967.

5. P.J. van Heerden, "Theory of optical information storage in solids," *Applied Optics*, pp. 393–400, Vol. 2, No. 4, 1963.

6. D. Psaltis, D. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," *Applied Optics*, pp. 1752–1759, Vol. 27, No. 9, 1988.

7. D.L. Staebler, W. Burke, W. Phillips, and J.J. Amodei, "Multiple storage and erasure of fixed holograms in Fe-doped $LiNbO_3$," *Applied Physics Letters*, pp. 182–184, Vol. 26, No. 4, 1975.

8. F. Mok, M. Tackitt, and H. M. Stoll, "Uniformly diffracting, angle multiplexed holograms in $LiNbO_3$," *OSA Annual Meeting 1989 Technical Digest Series*, p. 76, Vol. 18, Optical Society of America, Washington, DC, 1989.

a reprint from Applied Optics

Optical memory disks in optical information processing

Demetri Psaltis, Mark A. Neifeld, Alan Yamamura, and Seiji Kobayashi

We describe the use of optical memory disks as elements in optical information processing architectures. The optical disk is an optical memory device with a storage capacity approaching 10^{10} bits which is naturally suited to parallel access. We discuss optical disk characteristics which are important in optical computing systems such as contrast, diffraction efficiency, and phase uniformity. We describe techniques for holographic storage on optical disks and present reconstructions of several types of computer-generated holograms. Various optical information processing architectures are described for applications such as database retrieval, neural network implementation, and image correlation. Selected systems are experimentally demonstrated.
Key words: Optical memory disk, spatial light modulator, optical computing, computer-generated holograms.

I. Introduction

Ever since the introduction of the videodisc system in the late 1970s and the compact audio disk player in the mid-1980s, optical disk technology has been maturing at a rapid pace. Both write-once read-many (WORM) and magnetooptic read/write disk drives are presently available for high density storage on mainframes and personal computers. The conventional mode of both reading and writing used in present optical disk systems is serial. Specifically, a laser source will write one bit of data at a time on the disk, typically through a thermal mechanism.[1] Readout is achieved by using a lower power beam to illuminate the location of each bit on the disk individually and, based on the reflected or transmitted intensity detected, the bit is decoded as a logical 1 or 0. Although serial readout is well suited to conventional computers, the optical disk itself is naturally a parallel readout device.[2,3] To see this, consider illuminating a large portion of the disk with a collimated beam. The reflected or transmitted light contains all the data originally recorded in the illuminated area and a simple imaging system makes these data available to a detector array. This parallel access capability can be attractive when trying to solve memory access and contention problems in parallel computing architectures or when trying to implement an intelligent memory search procedure as with database machines.[4] Further, the optical disk represents a high resolution, computer controllable, spatial light modulator (SLM) which may be used in various optical computing architectures. For example, images stored on an optical disk may serve as a library of references in an optical image correlator and holograms stored on the disk may serve as interconnect patterns for hybrid optical/VLSI based neural networks.

In this paper we discuss the application of optical disk technology to areas in which parallel retrieval may be advantageous. We begin by characterizing the disk system used in our work, a Sony prototype sampled format drive with both WORM and magnetooptic media. In Sec. III we discuss parallel optical readout of 2-D blocks of data such as images. In the same section, we describe the use of optical disks as holographic storage media. We present and analyze several techniques for storing and retrieving data holographically and suggest some applications of holographic disk based storage. In Sec. IV we describe the use of optical disks as both storage and interconnect elements in neural network architectures. Finally, optical disk based image correlators are described and demonstrated in Sec. V. All the applications we discuss here are designed to combine the parallelism and interconnectivity of optics with a mature optical disk technology to result in feasible optical systems that perform useful computational tasks.

II. Characterization

The prototype Sony disk system used in most of our work (Fig. 1) can read and write both write-once and magnetooptic 5-in. reflective optical disks. The system records data as circular 1-μm diam pixels along a

Seiji Kobayashi is with Sony Corporation, Tokyo 100-31, Japan; the other authors are with the California Institute of Technology, Pasadena, California 91125.
Received 17 July 1989.
0003-6935/90/142038-20$02.00/0.

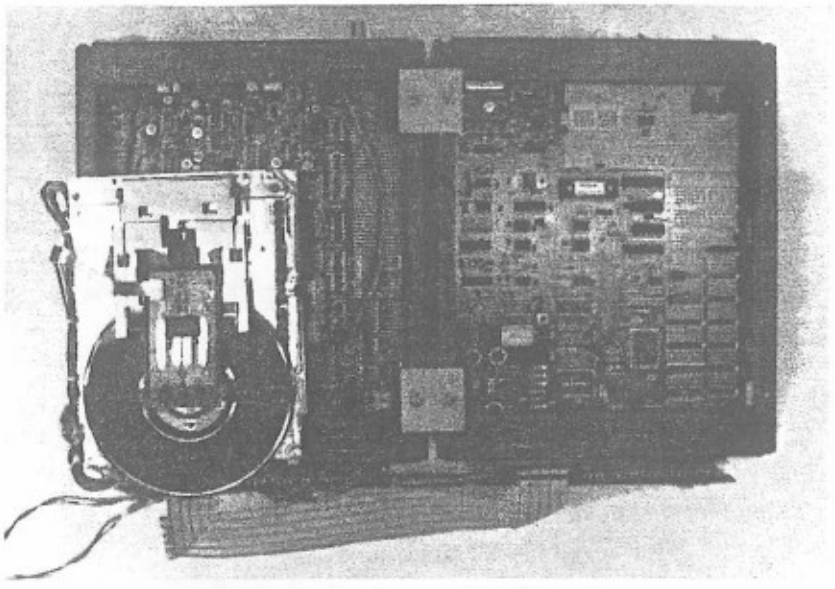


Fig. 1. Sony prototype optical disk system.

spiral on the disk with 20,000 turns between a 3-cm inner radius and 6-cm outer radius. The disk is divided into thirty-two sectors, and each loop of the spiral from the beginning of sector 0 to the end of sector 31 is called a track. Because the radius of the tracks changes gradually with angle, we often model the tracks as concentric circles separated by 1.5 μm. Pixels are recorded with a constant angular separation of .001°. This corresponds to an along track pixel-to-pixel separation that varies between .5 μm and 1 μm depending on radial position on the disk. This pixel recording density yields a storage capacity of over 7×10^9 bits on each side of the disk. The system is interfaced to a personal computer (PC) which provides serial read/write access to the disks. The system can read or write up to 15 million bits/s. Consequently a 1000×1000 image can be entered on the disk in 1/15 s. Since we can only write one line of the image per revolution, about 30 s are required to record the image in 2-D format on the disk. Note, however, that a thousand images using the same tracks could also be written during the same amount of time.

A variety of materials and recording mechanisms have been proposed for use in optical disks.[1] We briefly describe the recording mechanism employed in the write-once disk that we use in our experiments (see Ref. 5 for further details). The disk contains four thin metal alloy films of Sb_2Se_3 (300 Å), Bi_2Te_3 (150 Å), Sb_2Se_3 (1400 Å), and Al (1000 Å) formed by sputtering deposition on a glass or plastic substrate. The thickness of the various layers is chosen so that they form a low (5%) reflectivity interference filter. During the recording stage, a focused laser beam heats a spot of the Bi_2Te_3 layer through absorption. The Bi_2Te_3 and Sb_2Se_3 then form a four-element alloy by diffusion, eliminating the sharp interfaces between the layers. The low reflectivity interference filter is thus destroyed increasing the reflectivity of the medium to 12%. This reflectivity difference is detected during readout and decoded as a logical 1 or 0. The reflectivity of an interference filter is wavelength dependent. Our quoted figures are for the 633 nm He–Ne illumination that we use, but the thicknesses of the layers are chosen to maximize the change in reflectivity for the laser diode wavelength of 830 nm.

The magnetooptic disk contains a rare earth transition-metal alloy of TbFeCo. During the recording stage, the laser heats a spot on the disk above 180°C, the Curie temperature of the material. As the spot cools below this temperature, the material within the spot retains the magnetization of an external field applied perpendicular to the disk surface. The polarization of a low power readout laser rotates on reflection from the spot by an angle of ±0.15°, from the magnetooptic Kerr effect. The reflectivity of the magnetooptic disk is 17%, and the sign of the rotation angle depends on the direction of magnetization in the spot. This rotation is detected through a crossed polarizer and decoded as a logical 1 or 0. Depending on the setting of the polarizers, the amplitude of the light corresponding to the two states can be either on/off or plus/minus.[6]

Current disk systems use either continuous or sampled format schemes to maintain the position of the head over data in a track. Continuous format systems use a return signal either from a guide-groove embossed on the disk or the recorded data itself to constantly monitor and correct the position of the head relative to the data in a track. In contrast, sampled format systems, such as our experimental one, use tracking and timing information embossed along radial lines on the disk to periodically monitor and correct the head position. These lines of tracking and timing information appear every 270 pixels. Each line consists of a pattern of three embossed pits repeated in all 20,000 tracks as shown in Fig. 2. The first two pits provide tracking information. They are displaced an equal distance from the center of the track, one toward

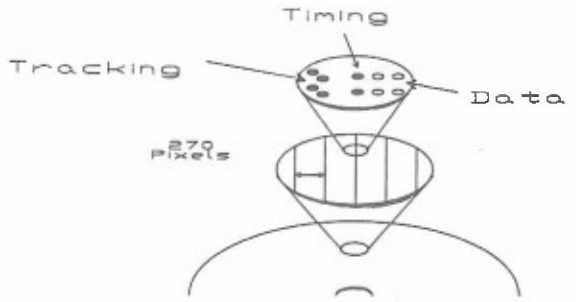
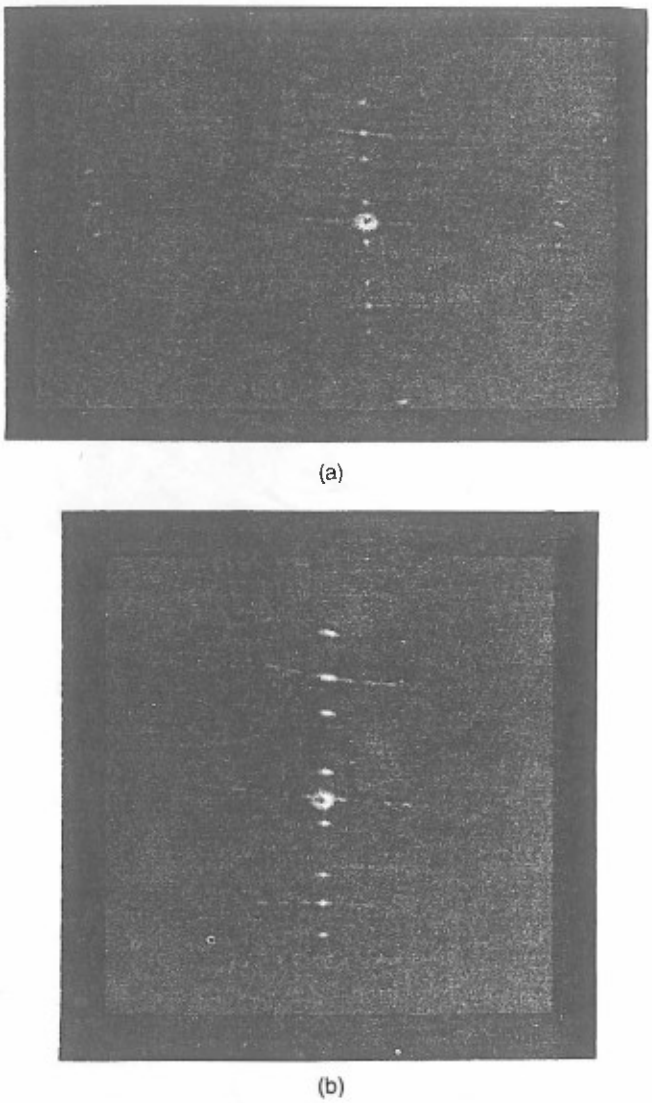


Fig. 2. Sampled-format tracking and timing information system.

the inside of the disk and the other toward the outside. If the head is exactly over the center of the track, the readout signal strength of the two pits will be equal; otherwise, the signal returning from one of the pits will be stronger than the other, thus indicating the direction to move the head. The third pit provides timing information used to synchronize the system clock and the disk.

There are two byproducts of the sampled format scheme that facilitate the parallel readout of data. First, the across track alignment of tracking and timing information combined with the synchronization between recorder and disk rotation allows us to specify the position of individual pixels with submicron accuracy in any one of more than one billion locations. This provides us with across track coherence, the ability to radially align pixels across different tracks. In contrast, with continuous format systems, the position of pixels in different tracks can drift by several pixel widths within a single sector. Second, the absence of guide-grooves allows us to retrieve high contrast images from low contrast media through schlieren imaging as described in Sec. III.

When we consider using the disk as a spatial light modulator, a number of additional performance issues arise. The resolution is determined by the track spacing in the radial direction (1.5 μm) and the minimum spot size in the azimuthal direction (0.5–1 μm). Notice that there is an inherent sampling in the radial direction due to the tracks. At the outer tracks, where the recorded pixels do not overlap, the image is also sampled in the azimuthal direction. We see later that we can make use of image diffraction caused by this sampling. Figure 3 shows the far field diffraction pattern when a grating recorded on the disk is illuminated. The grating was formed by periodically recording two tracks with all pixels on followed by two tracks with all pixels off. In Fig. 3(a), the grating is recorded on the inner tracks where pixels overlap along the track. In this case, the image is sampled in only one dimension thus producing diffraction orders in one dimension only. Figure 3(b) shows the diffraction with the same grating recorded at the outer edge of the disk. In this case, sampling in both dimensions results in a 2-D diffraction pattern. The maximum spatial frequency that can be recorded without aliasing of the image spectra is one-half of the sampling frequency in each direction.



(a)

(b)

Fig. 3. Far field diffraction pattern from grating: (a) grating written on inner tracks (R = 3 cm); and (b) grating written on outer tracks (R = 6 cm).

The reflectance function of the disk is basically binary both for the write-once and the magnetooptic disks. We have observed some dynamic range in the reflectivity of the write-once disks, controllable by varying the exposure for each pixel. We have not yet characterized fully the grey scale capability of the system. In any case, some form of area modulation can be used to encode multiple grey levels at the expense of space–bandwidth product. We will demonstrate one such method in the following section. The contrast of the light reflected from the disk is low for the Sony write-once disks (2:1). For magnetooptic disks, the polarization of the modulated light is orthogonal to the polarization of the incident light and the use of orthogonal polarizers in conjunction with the carrier encoding method discussed in the next section yields excellent contrast, limited primarily by the quality of the polarizers.

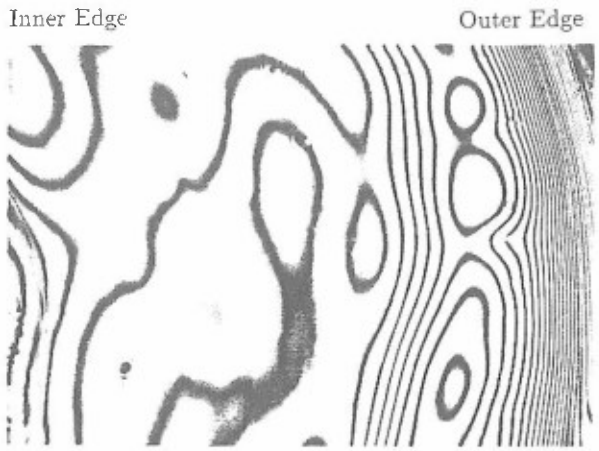


Fig. 4. Interferogram of Sony WORM disk.

Many of the processing architectures we propose use coherent processing techniques requiring phase uniformity across the surface of the disk. We have used a Fizeau interferometer to measure the phase uniformity of the Sony disks. Figure 4 shows a Fizeau interferogram of a 24- × 36-mm area of a glass-covered Sony write-once disk taken with a He–Ne laser source providing illumination. The figure shows numerous regions on the disk with optical thickness variations of less than a wavelength ($\lambda = 633$ nm) over distances of several millimeters. These regions are sufficiently large to contain images or holograms thousands of pixels on a side. The greater phase variation towards the outer edge of the disk is most likely caused by index variations due to stresses induced during manufacturing. We have also tested plastic-covered disks which generally show greater phase variation than the glass-covered ones.

In most uses, it would be more convenient if the optical disk system recorded pixels on a Cartesian grid. However, as noted earlier, our system actually writes pixels along curved tracks. We can neglect this curvature if we restrict attention to a small area of the disk. Consider a region at a distance R from the disk center. As shown in Fig. 5, we establish Cartesian coordinate axes with x in the azimuthal or along track direction and y in the radial or across track direction. Equation (1) converts the polar coordinates of the disk to the Cartesian coordinates in the region of interest:

$$\begin{Bmatrix} x = r\sin(\theta) \\ y = r\cos(\theta) - R \end{Bmatrix}. \tag{1}$$

The center-to-center spacing of the pixels in the radial dimension is δ_r and the angular separation between adjacent pixels is δ_θ in azimuth. We now superimpose a Cartesian grid on this pixel structure with x and y spacings as follows:

$$\begin{Bmatrix} \Delta_x = R\delta_\theta \\ \Delta_y = \delta_r \end{Bmatrix}. \tag{2}$$

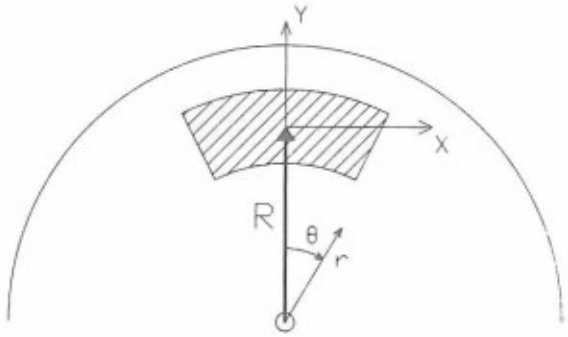


Fig. 5. Coordinate system for calculating the effect of track curvature.

This choice for Δ_x and Δ_y provides the best match between the pixels recorded on the disk and the points on the Cartesian grid. We now calculate the deviation of the pixel locations from their presumed Cartesian locations. The presumed coordinates of the points on the Cartesian grid are

$$\begin{Bmatrix} x' = n\Delta_x \\ y' = m\Delta_y \end{Bmatrix}, \tag{3}$$

whereas the actual location of the pixels on the polar grid are

$$\begin{Bmatrix} r = R + m\delta_r + n\dfrac{\delta_\theta\delta_r}{2\pi} \\ \theta = n\delta_\theta \end{Bmatrix}. \tag{4}$$

The actual Cartesian coordinates of the recorded pixels, therefore, are

$$\begin{Bmatrix} x = \left(R + m\delta_r + n\dfrac{\delta_\theta\delta_r}{2\pi}\right)\sin(n\delta_\theta) \approx n\Delta_x + nm\dfrac{\Delta_x\Delta_y}{R} \\ y = \left(R + m\delta_r + n\dfrac{\delta_\theta\delta_r}{2\pi}\right)\cos(n\delta_\theta) - R \approx m\Delta_y + n\dfrac{\Delta_x\Delta_y}{2\pi R} - n^2\dfrac{\Delta_x^2}{2R} \end{Bmatrix}. \tag{5}$$

We calculate the deviation between the actual pixel position and the presumed location on the Cartesian grid by subtracting Eq. (5) from Eq. (3):

$$\begin{Bmatrix} \epsilon_x = x' - x \approx nm\dfrac{\Delta_x\Delta_y}{R} = \dfrac{x'y'}{R} \\ \epsilon_y = y' - y \approx n\dfrac{\Delta_x\Delta_y}{2\pi R} - n^2\dfrac{\Delta_x^2}{2R} = x'\dfrac{\Delta_y}{2\pi R} - \dfrac{x'^2}{2R} \end{Bmatrix}. \tag{6}$$

For an array of 1000 × 1000 pixels on the Sony disks, the worst case pixel placement error is 1.25% of the array size (12.5 pixels) in the x-direction (at $R = 3$ cm, $\Delta_x = 0.5$ μm, and $\Delta_y = 1.5$ μm) and 0.14% of the array size (1.4 pixels) in the y-direction (at $R = 6$ cm, $\Delta_x = 1$ μm, and $\Delta_y = 1.5$ μm). In applications where this kind of positional error is not tolerable, we need to compensate the curvature through optical techniques and/or the recording geometry. We describe such methods in greater detail in the following section.

Diffraction efficiency is a key parameter in determining overall system efficiency since many of the

optical systems presented in the following sections use light diffracted from the disk. Given an accurate model of the surface reflectivity of the disk and how it will be used in an optical system, we can calculate the expected efficiency of the disk in that application. We model the reflectivity pattern of the disk using the following equation:

$$r(x,y) = r_0 + (r_1 - r_0)\left[\sum_{n,m} b_{nm}\delta(x - n\Delta_x, y - m\Delta_y)\right] \otimes s(x,y), \quad (7)$$

where b_{nm} represents the binary pixel pattern, r_0 and r_1 the reflectivity of unwritten and written pixels, respectively, $\otimes$ the convolution operation, and $s(x,y)$ the shape of each pixel.

The light reflected from the surface of the disk $E_r(x,y)$ is the product of the reflectivity pattern of Eq. (7) with the field of the illuminating light beam $E_i(x,y)$:

$$E_r(x,y) = E_i(x,y)r(x,y). \quad (8)$$

We can use Fresnel diffraction to calculate the field due to light reflected from the disk at any distance from the disk, as follows:

$$E(x,y,z) = \int\int E_r(x',y')\frac{z\exp(jkl)}{j\lambda l^2}dx'dy', \quad (9)$$

$$l^2 = (x - x')^2 + (y - y')^2 + z^2. \quad (10)$$

The efficiency η of the disk can then be found by integrating the intensity of light reflected by the disk over the region of space Σ where the optical system captures reflected light and dividing by the incident light energy:

$$\eta = \frac{\iiint_{\Sigma} |E(x,y,z)|^2 dxdydz}{\iint |E_i(x,y)|^2 dxdy}. \quad (11)$$

In the Appendix, as an example, we estimate the diffraction efficiency of the disk for schlieren imaging of the first diffracted order. Substituting parameters for the Sony write-once disk into Eq. (A7), [$|r_1|^2 = 0.12$, $|r_0|^2 = 0.05$, $\Delta_x = 0.5\ \mu$m, $\Delta_y = 1.5\ \mu$m, and $\Delta_r = 0.5\ \mu$m with $b(x,y) = 1$, $n = 0$, and $m = 1$], we find an estimated efficiency of $\eta_e = 0.112\%$ compared with a measured efficiency of $\eta_m = 0.114\%$. We estimate that the magnetooptic disk will be almost 1000× less efficient than the write-once disk in most applications. This large loss in efficiency was also observed experimentally.

III. Imaging and Holography

The fact that data can be retrieved in parallel from optical disks creates the possibility for eliminating some of the bottlenecks that currently exist in computers due to the mismatch between mass storage media and semiconductor memories.[2] A parallel random access memory would be one possible way to construct a parallel readout optical memory. In this case, M out of the N bits stored on a disk could be specified and retrieved simultaneously. In such a system, the apparatus that would scan the memory to realize this parallel retrieval capability would have to be set in $\binom{N}{M} \approx (N/M)^M$ distinct states to arbitrarily select any M-tuple. With $N = 10^{10}$ and $M = 1000$, we obtain about 10^{7000} distinct states. It is clearly not practical to realize an optical scanning mechanism that can do this. Therefore, we conclude that we must somehow structure the stored data to reduce the complexity of the access mechanism. The most straightforward way to impose such structure is to arrange the stored data in 2-D blocks, M bits each, that are retrievable in parallel. This reduces the number of choices the access mechanism addresses to a practical $\binom{N/M}{1} = 10^7$ for the previously quoted numbers. In this section, we discuss several methods, including holographic, for recording and retrieving 2-D blocks of data from optical disks.

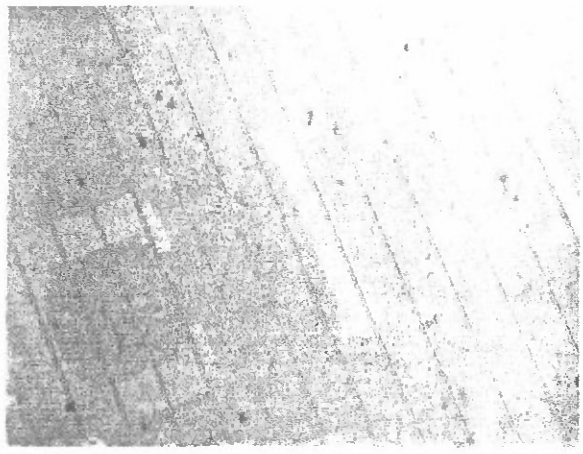


Fig. 6. Photograph of image written on Sony WORM disk.

Figure 6 is a photograph of a binary image written on the Sony write-once disk as viewed through a microscope. This image consists of 4024 × 512 pixels recorded on a polar grid. Note that the track curvature is not visible. The parallel lines, evident in the figure, are the radial strips of tracking and timing information described in the previous section. There are 270 pixels between each pair of these lines. Individual pixels are not discernible in this figure, but they are perfectly aligned in the radial direction resulting in the accurate recording of the letters in the figure. Note the poor contrast in Fig. 6. This is because the disk has an off-state (unwritten) reflectivity of 5%, while the on-state (written) reflectivity is only 12%. This large background and low differential reflectivity results in the poor contrast seen in the photo. Rilum and Tanguay used an interferometric technique to eliminate the background light obtained on reflection from a stamped optical disk.[7] This technique is not applicable to the Sony disk because the recorded data do not appear as embossed pits but rather as local variations in surface reflectivity. We use an alternative means for improving the contrast of the retrieved image. Since the binary image $b(x,y)$ to be recorded on the disk is sampled by a polar grid, light reflected from the disk will be diffracted into many orders or sidebands whose center frequencies will be determined by the

grid spacing. To make this clear, consider Eq. (12) for the reflectivity of the disk:

$$r(x,y) = r_0 + (r_1 - r_0)\left[b(x,y) \sum_{n,m} \delta(x - n\Delta_x, y - m\Delta_y) \right] \otimes s(x,y), \tag{12}$$

with the x-axis in the along track direction and the y-axis in the radial direction, as before. Since the background reflectivity r_0 is not sampled, energy in the first- and high-order diffracted fields arises only from the presence of the recorded image. Therefore, an image formed by the first-order diffracted field will not contain any bias light and will have high contrast. We can calculate how much energy is diffracted into the first-order and compare this to the total incident energy to obtain an estimate for the efficiency of the disk. The Fourier transform of the reflectivity function is

$$R(u,v) = r_0\delta(u,v) + (r_1 - r_0) \times \left[B(u,v) \otimes \frac{1}{\Delta_x\Delta_y} \sum_{n,m} \delta\left(u - \frac{n}{\Delta_x}, v - \frac{m}{\Delta_y} \right) \right] S(u,v), \tag{13}$$

where $B(u,v)$ and $S(u,v)$ are the Fourier transforms of $b(x,y)$ and $s(x,y)$, respectively.

The term that contributes to the formation of the desired image is

$$R_1(u,v) = \frac{r_1 - r_0}{\Delta_x\Delta_y} B\left(u, v - \frac{1}{\Delta_y} \right) S(u,v). \tag{14}$$

From these two equations we can express the disk efficiency as

$$\eta = \int\int |R_1(u,v)|^2 du dv. \tag{15}$$

For the Sony optical disk we have calculated the disk efficiency to be 0.112% (see Sec. II for parameters). This value agrees well with the measured efficiency of 0.114%. For a more detailed derivation of the efficiency of a schlieren imaging system used to image the optical disk surface, see the Appendix. An example of a high contrast image obtained by imaging the first diffracted order is shown in Fig. 7. The light diffracted by the tracks was selected to form this image.

As described in Sec. II, when we assume that pixels are written on a Cartesian grid, the presence of track curvature leads to positional errors given by Eq. (6). In the schlieren imaging system described above, the positional error of a recorded spot can lead to amplitude and phase errors in its contribution to the reflected field. We neglect the amplitude error since it only becomes significant when the position error $\bar{\epsilon}(x,y)$ is comparable to the distance between the disk and the image plane. The phase error observed in a given direction, however, can be approximated in most cases by the dot product between the wavevector in that direction $\bar{k}$ and the position error. Since $|\bar{k}|$ is large, this phase error cannot be neglected. When we consider only the first-order diffracted beam, $k_x = 0$ and the apparent phase error over the disk is given by

$$\phi_{\text{err}}(x,y;\bar{k}) = \bar{k} \cdot \epsilon(x,y) = -\frac{k_y x^2}{2R} = \frac{k}{2f} x^2, \tag{16}$$

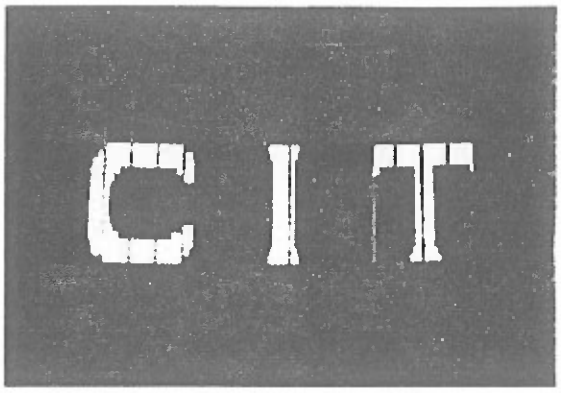


Fig. 7. High contrast image obtained by imaging the first-order diffracted component of light reflected by the disk used in Fig. 1.

$$f = -R\frac{k}{k_y} \tag{17}$$

$$k_y = \begin{cases} \dfrac{2\pi}{\Delta_y} \text{ for the +1 order,} \\ \dfrac{-2\pi}{\Delta_y} \text{ for the −1 order.} \end{cases} \tag{18}$$

This phase error can be modeled as a cylindrical lens at the disk plane with focal length $f = \pm R\Delta_y/\lambda$. For typical experimental parameters $R = 4.5$ cm, $\Delta_y = 1.5$ μm, and $\lambda = 633$ nm, the cylindrical focal length is 10.7 cm. This distortion can be corrected by an illumination system containing a cylindrical lens of focal length F. The product of the incident wavefront and the reflectance function of the disk in this case is

$$E_r(x,y) = E_i(x,y)\tilde{r}(x,y), \tag{19}$$

where $E_r(x,y)$ is the reflected field, $E_i(x,y) = \exp(j\alpha x^2)$ is the incident field corrected by the cylindrical lens, and $\tilde{r}(x,y)$ is the apparent reflectance function of the disk surface including the phase error. The illuminating optics should be chosen so that $E_r(x,y) = r(x,y)$, which yields $\alpha = -k/2f$. With this value for α, the incident illumination is given by $E_i = \exp(-jkx^2/2f)$ which can be generated by a line source located a distance f in front of the disk. A cylindrical lens with focal length F, at a distance $F + f$ in front of the disk, can be used to form the line source.

Correcting for this cylindrical lens effect does not, however, account for the positional errors of pixels due to the polar grid. In Sec. II we found that the position error of pixels in large images can exceed the pixel spacing. Since some applications require pixel position errors less than the interpixel separation, to minimize these errors we must make the interface to the disk conform to this polar recording format. For example, to accurately record an image sensed by a television camera, the camera should be modified to scan along curved lines matching the shape of the tracks on the disk.

Although the optical disk is basically a binary storage medium, it can also encode grey level images. Area modulation can be used to code multiple reflectivity

Fig. 8. Grey scale image written on optical disk using area modulation.

levels for superpixels consisting of several bits. For example, turning on n out of N pixels in a superpixel can be used to represent the integer value n. Various superpixel coding techniques have been investigated in the past.[8] We have implemented an area modulation scheme which uses a stochastic procedure to determine the position of on-pixel locations within each superpixel. In addition, this scheme improves the dynamic range of regions of low spatial frequency by stochastically selecting the value to be recorded in each element of an array of superpixels. Specifically, if a uniform region of p superpixels maps to a grey level between n and $n + 1$, randomly choosing between two levels for each superpixel in the region provides the entire p-superpixel region with an expected grey level equal to any one of $p - 1$ additional levels between n and $n + 1$. Figure 8 was generated using the area modulation recording scheme on the write-once disk and a schlieren imaging system as described above. The image shown consists of 512 × 480 superpixels each of dimension 6 × 4 pixels. The number of distinct grey levels recorded on the disk therefore was twenty-five. The grey levels present in the original image are clearly evident in the figure.

In addition to the recording and retrieval of images, the optical disk is an ideal medium for the storage of computer generated holograms (CGHs).[9–12,25] The imaging technique described above may be thought of as simply the reconstruction of an image plane hologram. Any other computer generated hologram can just as easily be stored on the optical disk. We have investigated various techniques for the calculation and recording of CGHs on the optical disk.

The first method we investigated is based on using the computer to form the holograms of the individual points that make up an image. Thresholding a Fresnel

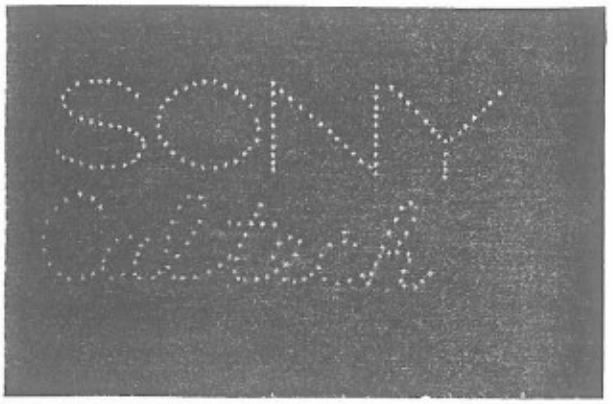


Fig. 9. Reconstruction of a binary Fresnel hologram stored on optical disk.

hologram of a single point (x_0,y_0,z_0) yields a Fresnel zone plate. It can be calculated by simply evaluating the real part of the Fresnel transform of a single point and thresholding the result, as shown below:

$$h_p(x,y) \propto \operatorname{sgn}\left[\operatorname{Re}\left\{\int\int \delta(x' - x_0, y' - y_0) \times \exp\left(j\frac{\pi}{\lambda z_0}\right)[(x' - x)^2 + (y' - y)^2]dx'dy'\right\}\right], \quad (20)$$

$$\propto \operatorname{sgn}\left[\cos\left(\frac{\pi}{\lambda z_0}[(x - x_0)^2 + (y - y_0)^2]\right)\right]. \quad (21)$$

The reconstruction of the object point is achieved by illuminating the reflection hologram $h_p(x,y)$ with a plane wave. One component of the reflected field is a spherical wave converging to the point (x_0,y_0,z_0). In a similar fashion, the Fresnel hologram of multiple points is calculated by summing many individual holograms and thresholding the result:

$$h(x,y) = \operatorname{sgn}\left[\sum_m \cos\left(\frac{\pi}{\lambda z_m}[(x - x_m)^2 + (y - y_m)^2]\right)\right], \quad (22)$$

where (x_m,y_m,z_m) are the coordinates of the dots with which we construct the image that is stored holographically. The y_m terms are chosen in the range from 0 to $\lambda z_0/\sqrt{\Delta_y^2 - \lambda^2}$, and the x_m terms in the range from 0 to $\lambda z_0/\sqrt{\Delta_x^2 - \lambda^2}$. This guarantees that the reconstruction of the first-order in the y-direction does not overlap with any of the other orders. In effect, this is how we construct an off-axis hologram. If the number of dots that comprise the image is far less than the space-bandwidth product of the hologram, a reconstruction with high signal-to-noise ratio (SNR) is obtained.

Figure 9 shows the reconstruction of a binary Fresnel hologram we have recorded on a write-once disk. The reconstructed object contains over 100 points. The hologram consists of a 1000 × 1000 array of pixels recorded on the disk at a radius equal to 4 cm. By simply illuminating the hologram with a raw beam, we observe a self-focused reconstruction at a distance of 10 cm from the disk. The size of the reconstructed image is 3 × 5 cm which is much larger than the size of

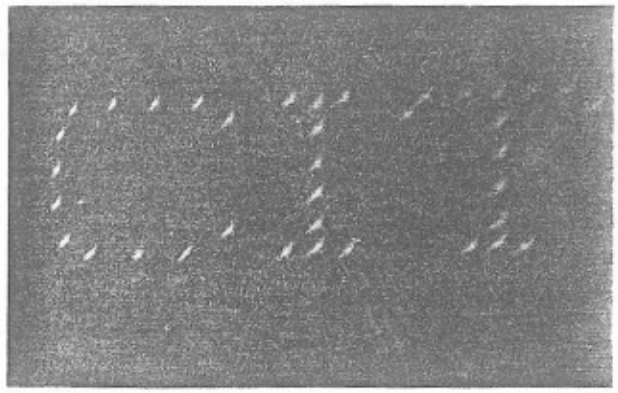


Fig. 10. Reconstruction of Fresnel hologram stored on magnetooptic disk.

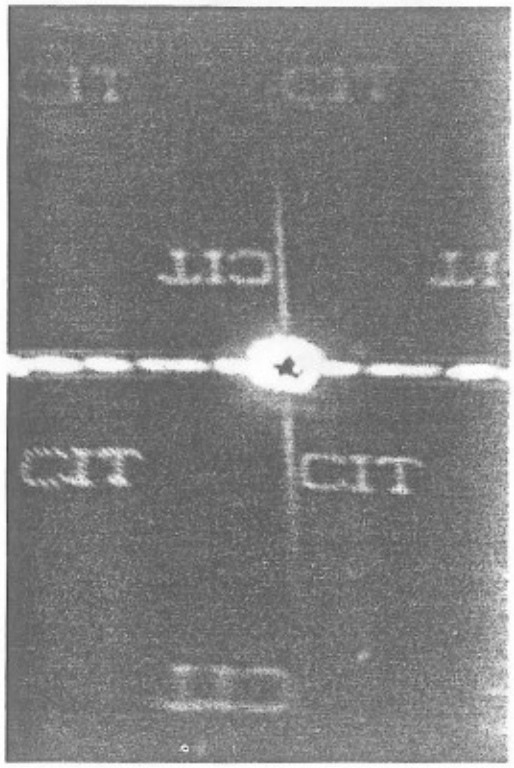


Fig. 12. Reconstruction of Fourier transform hologram.

the hologram itself (0.7 × 1 mm). This is a convenient feature which allows us to match the high density storage medium with a lower resolution array of detectors. Figure 10 shows the reconstruction of a Fresnel hologram recorded on the magnetooptic disk. In this case, the reconstructed hologram was observed through crossed polarizers to enhance the contrast. This polarization filtering technique, made possible by the polarization switching property of magnetooptic modulation, is necessary in this case because of the lower efficiency of the magnetooptic disk.

We can also record any conventional CGH (e.g., Lohmann or Lee).[13-15] One CGH technique we have investigated consists of forming 1-D superpixels with position encoding of phase information. As can be seen in Fig. 11(a), each 1-D superpixel consists of four pixels along a specific track. By selecting the appropriate combinations of pixels to record within a group, each superpixel can be made to represent any of the nine phasors including the zero phaser shown in Fig. 11(b). We may also be able to represent a larger number of phasors without increasing the size of the superpixels by varying the size of the recorded pixels. In our system, this can be accomplished by either changing the energy used to write each pixel or by multiply exposing pixels.

We can also store Fourier holograms by simply recording the complex Fourier transform of the object field as described above. The Fourier transform hologram can be calculated in a number of ways; for example, we could use the fast Fourier transform algorithm or we could calculate a Fresnel hologram using Eq. (22) with a large z_0. This latter approach is equivalent to calculating the Fraunhofer diffraction pattern of the object and is the approach we have chosen. Shown in Fig. 12 is the reconstruction from a binary Fourier transform hologram. The data for this hologram were computed by thresholding the real part of the Fourier transform of the object. The hologram consists of 1000 × 1000 pixels written on the disk at a radius of 4 cm. Reconstruction was achieved using a Fourier transform lens.

One can also combine the Fourier and Fresnel hologram methods described above. For example, Fig. 13 shows the image reconstructed from a hologram generated using the Fourier transform in the along track direction and the Fresnel hologram in the across track direction. This transformation is given by the following equation:

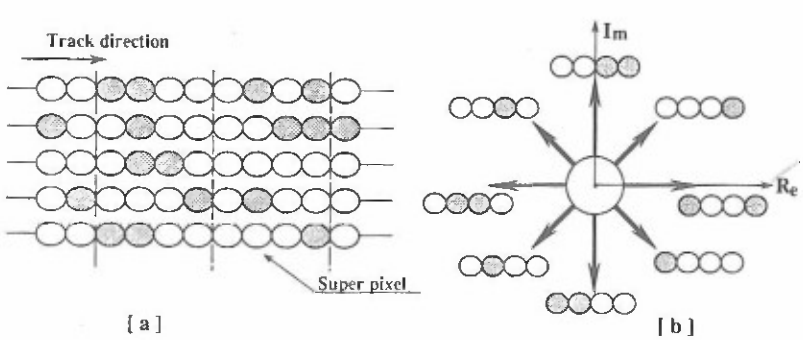


Fig. 11. Schematic of CGH technique used to record complex valued pixels: (a) one-dimensional superpixel used to code phase information; and (b) complex values recordable using (a).

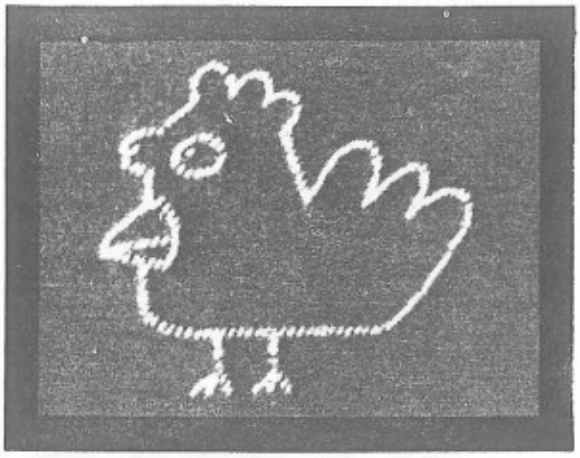


Fig. 13. Reconstruction of Fourier along track/Fresnel across track hologram.

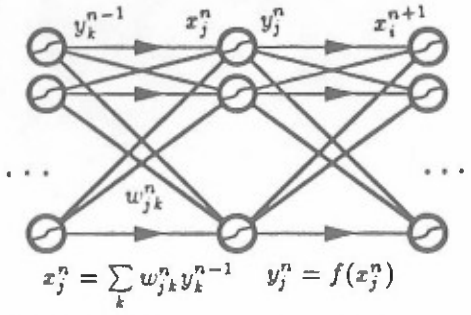


Fig. 14. Multilayer feedforward neural network.

$$h(x,y) = \int\int b(x',y') \exp\left[j\frac{\pi}{\lambda z_0}(y-y')^2\right] \exp\left(-j\frac{2\pi}{\lambda z_0} xx'\right) dx'dy'. \tag{23}$$

As before, we record a thresholded version of the real part of $h(x,y)$. Reconstruction is achieved using a cylindrical lens. Another storage scheme might involve calculating the 1-D Fourier transform of each line of the object and storing these 1-D signals along separate tracks. One benefit of using the Fourier transform in the along track direction and an imaging lens across track is that across track coherence is no longer required for parallel readout of 2-D data arrays since each track reconstructs a shift invariant line of data.[16]

We would also like to use the shift invariance property of the Fourier holograms in the along track direction to generate stationary reconstructions from a rotating disk. Unfortunately, this is not exactly true. As the disk rotates, the hologram experiences both translation and rotation. Rotation of a Fourier transform hologram results in an equal rotation of the reconstruction around the axis defined by the direction of the zero-order reflected beam. Thus, disk rotation results in apparent rotation and translation of the reconstructed image. We can neglect this motion if the reconstructed image moves less than a resolution element during the observation period. The size of a resolution element in the reconstruction from an $N_x \times N_y$-pixel Fourier transform hologram is given by

$$\begin{Bmatrix} \delta_x = \dfrac{\lambda F}{\Delta_x N_x} \\ \delta_y = \dfrac{\lambda F}{\Delta_y N_y} \end{Bmatrix} \tag{24}$$

in the x- and y-directions, respectively. F is the focal length of the Fourier transform lens used in the reconstruction.

The shift observed in the reconstruction plane as a function of disk rotation angle θ is

$$\begin{Bmatrix} \varepsilon_x(\theta) = \dfrac{\lambda F}{2\Delta_y}[\cos(\theta) - 1] \\ \varepsilon_y(\theta) = \dfrac{\lambda F}{2\Delta_x}\sin(\theta) \end{Bmatrix}. \tag{25}$$

The hologram subtends an angle $\theta = N_x\delta_\theta$. In order that the reconstruction remain stationary during a rotation of the disk by θ, we find the following limits on the size of the hologram:

$$\begin{Bmatrix} N_x < \dfrac{\lambda F}{\Delta_x \varepsilon_x(\theta)} \\ N_y < \dfrac{\lambda F}{\Delta_y \varepsilon_y(\theta)} \end{Bmatrix}. \tag{26}$$

Making a small θ approximation, we arrive at the following set of constraints for N_x and N_y:

$$\begin{Bmatrix} N_x < \dfrac{\sqrt[3]{4\Delta_y/R}}{\delta_\theta} \\ N_x N_y < \dfrac{2R}{\Delta_y} \end{Bmatrix}. \tag{27}$$

Thus, for the disk parameters in our experiment, the holograms must be smaller than 300 × 300 pixels, if we require the motion of the reconstruction to be negligible.

IV. Neural Networks

Neural network architectures are particularly well suited for the use of optical disks. This is due to the large memory that is typically required for the storage of the interconnecting weights in large networks. The optical disk not only supplies this storage capability, but in addition it provides the rapid access that is necessary for fast computation of the mapping of the network. Figure 14 shows a multilayer feedforward network, the most common neural network architecture. In this architecture, neurons are grouped in layers. The input to the jth neuron in the lth layer is x_j^l and its output is y_j^l. The weight of the interconnection between the input of this neuron and the output of the kth neuron in the previous layer is w_{jk}^l. Equation (28) governs the operation of the network:

$$\begin{Bmatrix} x_j^l = \sum\limits_k w_{jk}^l y_k^{l-1} \\ y_j^l = f(x_j^l) \end{Bmatrix}, \tag{28}$$

where f is typically a sigmoidal function.

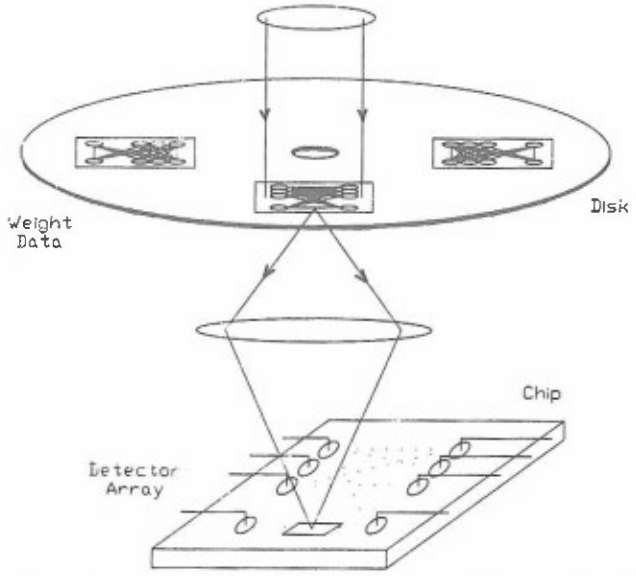


Fig. 15. Optical disk/VLSI hybrid neural network implementation.

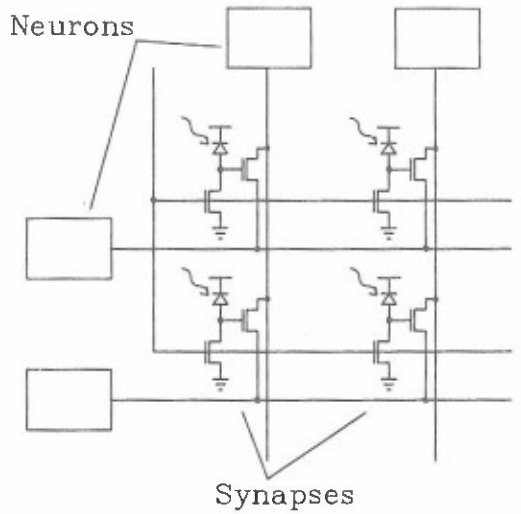


Fig. 16. Schematic of VLSI neural network crossbar implementation.

An implementation of a feedforward network using optical disks combined with analog VLSI chips is shown in Fig. 15. The VLSI chip implements a single layer of the network. The weights that are used for propagating through one layer are optically loaded via the third dimension from the disk. An array of photodiodes is integrated into the VLSI chip for this purpose. The weights for the different layers of the network are stored adjacent to one another along the azimuthal direction of the disk. To achieve a multilayer network, we first download the weights for the first layer. The VLSI chip then evaluates the response of the first layer and stores the result. Meanwhile, the disk is spun so that the weights of the next layer are aligned with the chip, the new weights are downloaded and the response for the next layer is evaluated on the chip. This procedure is repeated until the response of the final layer is evaluated.

There are several possible implementations of the basic idea described in the previous paragraph. We describe a particular chip design which we are implementing experimentally. A schematic diagram of this chip is shown in Fig. 16. The neurons and synapses are arranged as a crossbar. The output of each neuron in the top row is a voltage source that raises the potential of its corresponding vertical wire to 5 V if the neuron is on or sets it to 0 V if the neuron is off. The amount of current that flows from each vertical wire into a horizontal wire is determined by the channel resistance of the field effect transister FET at the corresponding intersection. The total current in each horizontal wire is the sum of the currents that were contributed from all the vertical wires. This summed current becomes the input to the neuron attached to each horizontal line. The neuron circuit accepts the input current, thresholds it, and generates an output voltage. The entire circuit is actually bidirectional. A horizontal wire that was previously used to sum the currents that flow from the vertical wires, can also be used to broadcast the state (voltage) of the neuron attached to it. Consequently, if the weights can be dynamically updated, a single chip can be used for the implementation of a multilayer network with data going back and forth between the two sets of neurons.

The strength of the connection between two neurons is increased by raising the gate voltage of the FET located at the intersection between the corresponding wires. The gate voltage is controlled by a circuit consisting of a series combination of a reverse-biased photodiode and a second, reset transistor. At the beginning of each cycle, the reset transistors are turned on, which sets the gates of the synapse transistors to 0 V. The reset transistors are then turned off for the rest of the cycle. The weights that are stored on the disk are imaged onto the chip. To facilitate this description, we will assume for the moment that the weights are binary (0,1) and we discuss later in more detail how to handle multivalued weights. When light from the disk strikes a photodiode on the chip, current flows through the photodiode charging the gate of the synapse transistor. This turns the transistor on, allowing current to flow between horizontal and vertical wires. We have not yet completed the testing of this chip, but we have characterized a circuit that consists of just two synapses that are however much more complex.

Figure 17 is a circuit diagram of one of the synapses that were fabricated. This synapse is a multiplying digital-to-analog converter (MDAC)[17] contributing a current to the target neuron proportional to the product between the signal received from the input neuron and the weight, encoded as a binary quintuple, that is optically received by a set of five photodiodes. The weight and the neuron activity can be bipolar. This is achieved through the use of a pair of horizontal and vertical wires for each neuron. We denote the wires in the vertical and horizontal directions by V_m^+, V_m^- and

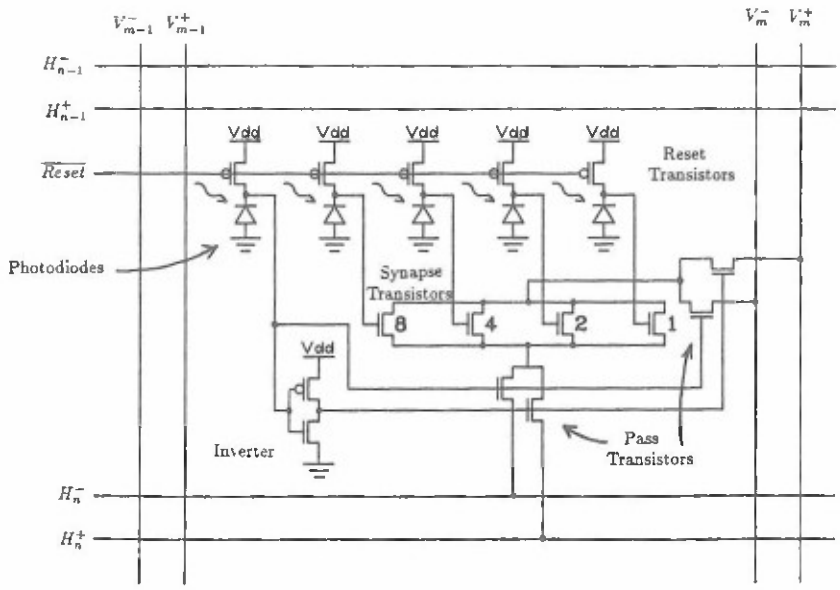


Fig. 17. Synapse circuit diagram.

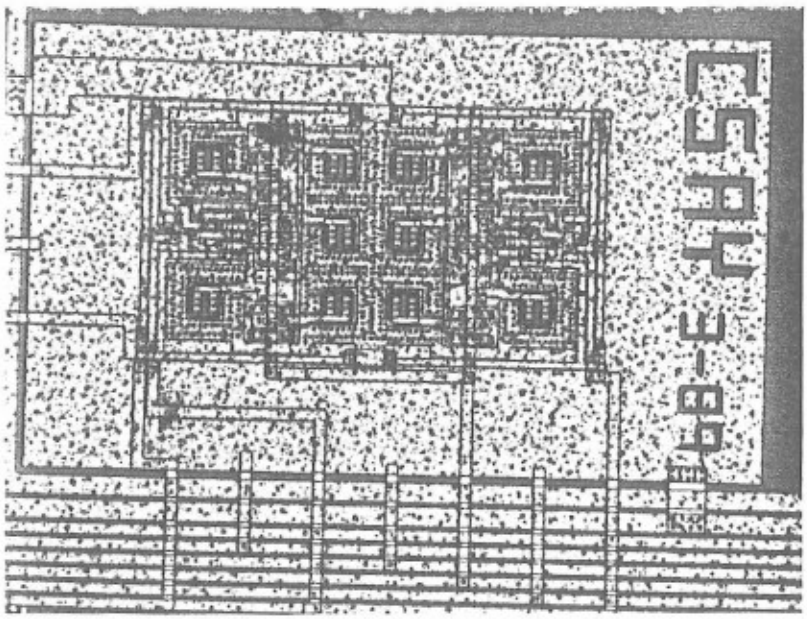


Fig. 18. Photograph of MDAC synapses.

H_n^+,H_n^-, respectively. Four of the five photodiodes are dedicated to receiving the four bit weight value and the fifth, on the far left of Fig. 17, is for the sign bit. The five reset transistors in this design initially charge the gate voltages of the synapse transistors to V_{dd}. There are two pairs of pass transistors shown in Fig. 17. At any one time, one transistor of each pair is on and the other is off. In this manner, the setting of the pass transistors determines which two perpendicular wires are connected. The setting of the pass transistors is controlled by the sign bit with the help of the inverter circuit. The strength of the interconnection is determined by the setting of the synapse transistor, which in turn is determined by the optical signal on the photodiodes (the binary encoded weight). The channel resistance of an FET is proportional to its width-to-length ratio (W/L). The analog-to-digital conversion is performed by scaling logarithmically W/L in the four synapse transistors. For example, suppose we are receiving a positive input signal along the vertical wires and the weight is negative. In this case, both vertical wires are set to 5 V by the output stage of the associated neuron. The sign bit through the pass transistors closes the circuit between the V_m^- and H_n^- wires. The amount of current that flows to H_n^- is proportional to the integer value corresponding to the four bit binary number detected by the photodiodes and also depends on the voltage setting ($V_{ref} \leq 5$ V) of H_n^-. If the sign bit flips, current will flow in the same direction but on H_n^+. The target neuron subtracts the currents on H_n^+ and H_n^- to form its output and hence a bipolar weight is realized.

The photograph in Fig. 18 shows the pair of synapses which was fabricated through MOSIS. The circuitry for the synapses is protected by a layer of metal. Each of the ten squares in the figure containing π shaped features is a 26- × 27-μm opening in this protective layer which allows light to strike the 20- × 21-μm photodiode located underneath. The plot of Fig. 19 shows the current on H_n^+ vs V_{ref} for voltages of 5 V (neuron on) and 0 V (neuron off) on V_m^+. When V_{ref} is set to 0 V (5 V), the resulting current swings between 0 μA (−140 μA) and 140 μA (0 μA). In other words, when V_{ref} is set to either 0 V or 5 V, this is a two-quadrant multiplier. If, however, V_{ref} is set to any intermediate value, the output current swing is bipolar. In particular, for the circuit we fabricated, when V_{ref} = 1.17 V, the positive and negative deviations of the current are symmetric and we have a balanced four-quadrant multiplier.

The speed of the neural network chip is limited principally by the time required to discharge the gates of the synapse transistors through the photodiodes. The discharge current depends on the amount of light striking it and the voltage across the photodiode. Figure 20 shows the current flowing though the synapse as a function of time as all the photodiodes are simultaneously struck with light for three different intensity levels; V_{ref} was set to 1.17 V for this measurement. As expected, the response time is inversely proportional to the light intensity and corresponds to a switching energy of 1.55 pJ. This chip was not optimized for speed (phototransistors could be used in place of the photodiodes); however, we can use the measurement of Fig. 20 to obtain an estimate for the speed with which we can operate such a network. Let us assume that we have available optical power equal to 1 W and a chip consisting of 100 × 100 synapses. The diffraction efficiency of the disk was estimated to be approximately 0.1% which yields 1 mW total power incident on the chip, or 100 nW per synapse. Dividing the switching energy by the available power per synapse, we obtain 15.5 μs response time. This corresponds to 3 Gbits/s transfer rate between memory and chip. Even though this is a remarkable rate, there is a lot of room for improvement in the speed in this design, through a disk with better light efficiency, the use of phototransistors, and a reduction in the number of bits used in the MDACs.

In the neural network implementation described above, the disk is used only as a parallel read-out storage device and the electronic chip is used to perform all the calculations. It is also possible to use the disks as computer-generated holograms or transparencies in many of the optical neural network architectures that have been previously proposed.[18] In such optical neural networks, the analog multiplication needed to implement the weights is performed by propagating an incident light field through a transparency (the disk in this case) and summing multiple such products onto a single detector location. To accomplish this, the output of each neuron must be an optical signal. This can be done through the use of spatial

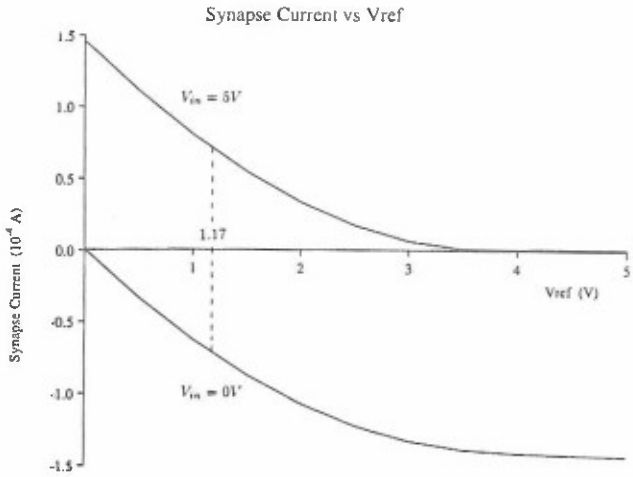


Fig. 19. Synapse current dependence on V_{ref}.

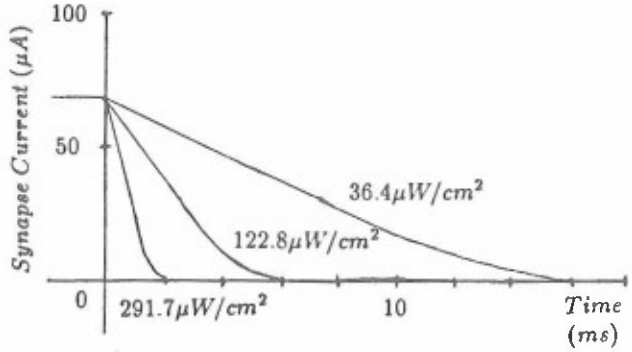


Fig. 20. Synaptic time response dependence on light intensity.

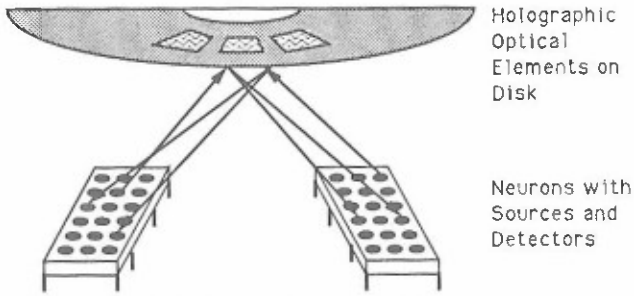


Fig. 21. Optical disk implementation of synaptic connections.

light modulators or optoelectronics. The optoelectronic approach for building the neurons combined with implementation of the synapses on the disk is schematically shown in Fig. 21. The neurons can be fabricated in GaAs[19] on which detectors, sources, and electronic circuits can be monolithically integrated. We can greatly increase the density of neurons on the GaAs chip compared to the previous approach, since we no longer require circuitry on the chip to detect the weights and multiply them with the neuron outputs. Second, since optics provides us with greater flexibility in performing the interconnections between the neu-

rons, we can, at least in principle, construct not only larger but also more general neural network architectures beyond multilayer feedforward networks.

V. Correlators

The correlation function $c(\tilde{x},\tilde{y})$ is defined as

$$c(\tilde{x},\tilde{y}) = \mathcal{F}^{-1}\{F(w_x,w_y)G^*(w_x,w_y)\} \tag{29}$$

$$= \int\int f(x,y)g(x-\tilde{x},y-\tilde{y})dxdy, \tag{30}$$

where $f(x,y)$ and $g(x,y)$ are two real images, $F(w_x,w_y)$ and $G(w_x,w_y)$ are their respective 2-D Fourier transforms, and $\mathcal{F}^{-1}\{\}$ is the 2-D inverse Fourier transform operator. It is well known that $c(\tilde{x},\tilde{y})$ is sharply peaked at the point (x_0,y_0) when $f(x,y) = g(x-x_0,y-y_0)$. This property is what makes the correlation function useful for pattern recognition because, regardless of the position of the input image $g(x,y)$, $c(\tilde{x},\tilde{y})$ will have a peak if $f(x,y)$ and $g(x,y)$ are matched. Since, in general, there are many versions of an image g that we would like to recognize, a reliable image recognition system should provide invariance to multiple object attributes. Often the best way to achieve this invariance is to use a large number of reference patterns f against which to compare g in order to obtain reliable recognition.

Optical image correlators based on Fourier transform (FT) holograms were proposed by VanderLugt in 1964.[20] For optical correlation to be a realistic approach to image recognition, we require a memory device sufficient to store a large reference image library, an SLM which interfaces with this memory in real time, and a scanning or addressing mechanism which allows interrogation of the entire reference library in a reasonable amount of time. The optical disk provides these three characteristics in one device. In this section, we describe several optical disk based image correlation architectures and present experimental results taken from selected systems.[21,22] We will examine critical parameters associated with each architecture and evaluate each system in terms of power and speed.

A. VanderLugt CGH Correlator

The first disk based image correlator to be described is the simple VanderLugt correlator shown in Fig. 22. As can be seen from the figure, a Fourier transform computer-generated hologram recorded on the optical disk is used as a Fourier plane filter for the input image. The product of the transforms of the input and reference images is formed at the disk and an inverse transform yields the desired 2-D correlation in the output plane. As the disk rotates, a new correlation pattern is generated every time a different CGH aligns with the input image FT. Therefore, whenever there is a match between the input FT and the CGH, a peak occurs in the output plane of the system. The location of this peak, which corresponds to the location of the object of interest in the input plane, may be anywhere in the correlation plane. Therefore, a 2-D detector array is required to acquire the correlation data. Furthermore, since the correct correlation only occurs during the brief periods of alignment between the input FT and the CGH, the detector array need only be queried at these times. A pulsed laser or an electronically gated detector array could be used to achieve the appropriate sampling. The proper operation of this system depends on the optical quality of various disk coating materials which are capable of introducing random phase distortions across the system filter plane. As we discussed in Sec. II, however, the optical quality of several commercially available disks is sufficient to make them suitable for these applications.

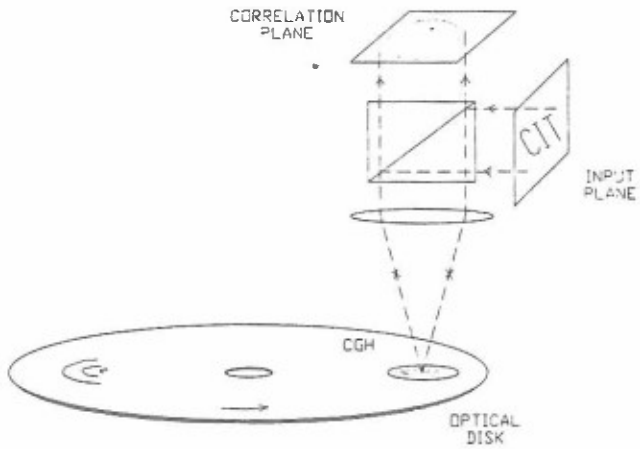


Fig. 22. Optical disk based VanderLugt correlator.

As with other FT based architectures, one advantage of this correlator is its potentially high speed. The correlation rate in this system is limited by disk rotation rate and detectability of the correlation peak. Taking a typical disk rotation rate of 40 Hz, we calculate a correlation rate R_c = 400,000 correlations/s for 100 × 100 pixel images. This correlation rate implies that to detect a correlation peak, each element of the 2-D detector array in the output plane must have a bandwidth of ≈ 400 kHz. Further, we can calculate the peak detectibility n_p given by the number of photons detected at the correlation peak, by first calculating the peak dwell time τ, multiplying this by the expected power in the correlation peak P_c, and dividing by the photon energy. That is:

$$n_p = \tau P_c/h\nu, \tag{31}$$

where ν is the frequency of the optical field. Peak dwell time is given simply by $\tau = 1/NR_c$ where N is the number of pixels in the reference image in the along track direction. In the case of random, bipolar images, the expected fraction of diffracted power that will be measured at the peak is 1/2. By random we mean that each pixel of the image is equally likely to take on either of the two possible values ±1. Given the disk diffraction efficiency, η, and the source power P_s, the correlation peak power is

$$P_c = \eta P_s/2. \tag{32}$$

Substituting P_s = 10 mW and η = 0.1%, we find an expected peak power P_c = 5 μW. This rate leads to an

easily detectable 10^5 photons in the correlation peak. The two most significant drawbacks of this system are alignment criticality and computational overhead. For each reference filter, a 2-D FT CGH must be computed and written on the disk. For a large reference library the time required for this procedure can be long. More importantly, the alignment of the input FT and the CGH is critical to within the resolution of the CGH ($\approx 1\ \mu$m). As the disk rotates, nonuniformities resulting from wobble and disk center offset, lead to nonuniformities in reference image location with respect to the optical system. For example, the Sony WORM disks we use allow up to 1° of wobble and up to 50 μm offset between the rotational center of the disk and the actual track center. These nonuniformities must be compensated for the output correlation to be accurate.

B. Photorefractive Correlator

Since the wobble and offset problems introduce slowly varying nonuniformities (<50 wobble cycles/rotation), the problem of alignment sensitivity can be effectively dealt with using real time compensation with feedback of the sort used in commercial disk drives; however, the computational overhead associated with generating the desired reference library in the above system remains a problem. The system of Figure 23 eliminates this processing time by allowing the reference images themselves to be recorded on the disk instead of FT CGHs. In this system, a photorefractive crystal or any other real time, temporary holographic storage medium is used to record a hologram of the input FT. During the recording phase, the disk illumination is blocked and the input transparency is illuminated from the right. A hologram is formed between the input FT and the reference beam as shown. This hologram will then be read out using the reference library. On readout, the input is blocked and the disk is illuminated. The product of the input and reference FTs is formed in the crystal and inverse transformed to yield the correlation output.

If the photorefractive crystal is replaced by a thin medium such as a holographic plate, then the output pattern is exactly the desired 2-D correlation; however, it has been shown that when a thick hologram is used in the filter plane of such a system, the resulting output is a 1-D slice of the 2-D correlation pattern.[23] This can be understood by considering the recording arrangement shown in Fig. 23. On recording, each plane wave corresponding to one of the points in the input image forms a grating with the reference beam. The resulting hologram exhibits Bragg selectivity in the horizontal direction. On readout, a point along a given radial line on the disk can only read out those gratings formed by points along one vertical line in the input. Each such line on the disk reads out a corresponding array of holograms and generates a vertical array of spots in the correlation plane at the horizontal location corresponding to the reference beam FT. The coherent sum of all such reconstructions comprises the output of the correlator. This output pattern is the desired 2-D

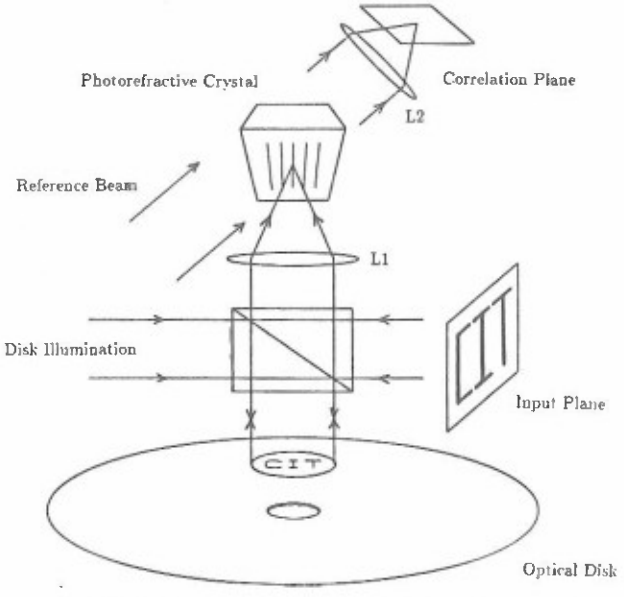


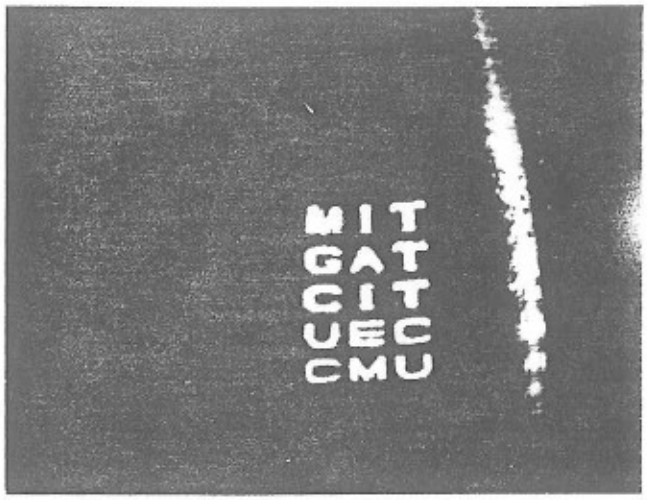
Fig. 23. Photorefractive/optical disk based correlator.

correlation multiplied, in the horizontal direction, by a sinc function whose width is inversely proportional to the hologram thickness. In the system of Fig. 23, this property does not cause problems since all 1-D slices are obtained sequentially via disk rotation. Further, instead of requiring a full 2-D detector array at the output, a 1-D array is sufficient to sequentially detect each slice of correlation output. Despite the advantages gained in terms of computational overhead and detector simplicity, alignment compensation remains a critical issue with this system. The expected correlation rate obtainable using this system is again limited primarily by disk speed and peak detectability. A rate of 400,000 correlations/s is still easily achievable, yielding a detected signal at the correlation peak of more than 10^5 photons using again a 10 mW source.

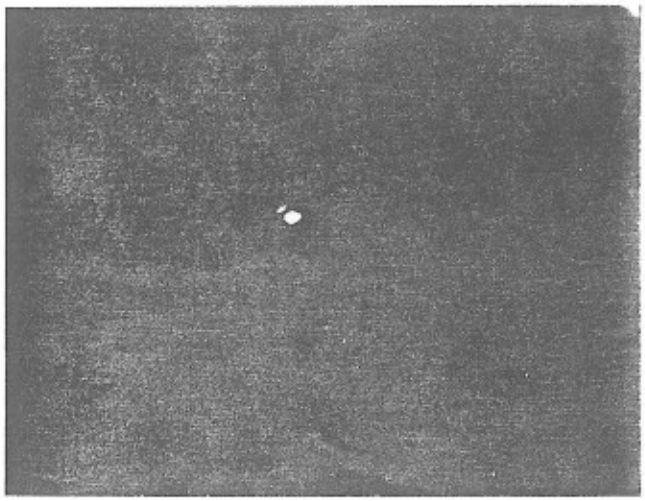
This system has been experimentally demonstrated using a thin hologram in place of the photorefractive crystal. For this experiment, we chose to use a Sony WORM disk as an SLM at the input as well as the reference. The results are shown in Fig. 24. Figure 24(a) shows the image recorded on two Sony disks; one disk was used to record the filter plane hologram on the plate and the another was used to read out the hologram. The correlation plane output is shown in Fig. 24(b). The characteristic autocorrelation peak appears in the output.

C. Rotating Mirror Correlator

The previous two Fourier transform based correlators implement Eq. (29) to generate the correlation function. While FT based systems are typically superior in terms of speed, alignment and coherence requirements are relative disadvantages. We discuss next two systems which perform 2-D correlations based on Eq. (30). In these systems the correlation function is generated by calculating an inner product for every relative shift between input and reference



(a)



(b)

Fig. 24. Disk based Fourier plane correlator results using a plate: (a) input image recorded on disk 1 and used to record the hologram; (b) correlation pattern obtained using disk 2 as reference to read out the hologram.

images. Since these shifts will be generated sequentially, the correlation will appear as a 1-D signal representing a raster version of the desired 2-D correlation pattern. As we will see, these systems sacrifice correlation rate for operational simplicity without alignment criticality, while at the same time relieving source coherence requirements.

A simple image plane correlator is shown in Fig. 25. An image of the input scene is formed at the disk on which a library of reference images resides. The total transmitted or reflected light is collected by a detector at the output. The rotation of a polygon mirror causes the input image to scan the disk radially while the disk rotation itself provides scanning in the orthogonal direction. The detected light therefore represents the instantaneous inner produce between the input and a shifted version of one of the references. All relative shifts between input and reference images are generated using these two scanning mechanisms and as a result, the light collected at the output is exactly a 1-D raster signal of the desired 2-D correlation. An example of the output obtained from the rotating mirror correlator is shown in Fig. 26 along with a computer simulation of the desired 2-D autocorrelation function displayed as a 1-D raster. Figure 26(a) is the reference image written on a Sony write once disk. This image has up to 6912 pixels along track and comprises 1024 tracks. The input to the system was provided by a transparency of the acronym CIT illuminated by a He–Ne laser. Figure 26(c) shows the correlation signal generated by the optical system for this input. The asymmetry in the optical autocorrelation is due to a slight mismatch between the input and reference images.

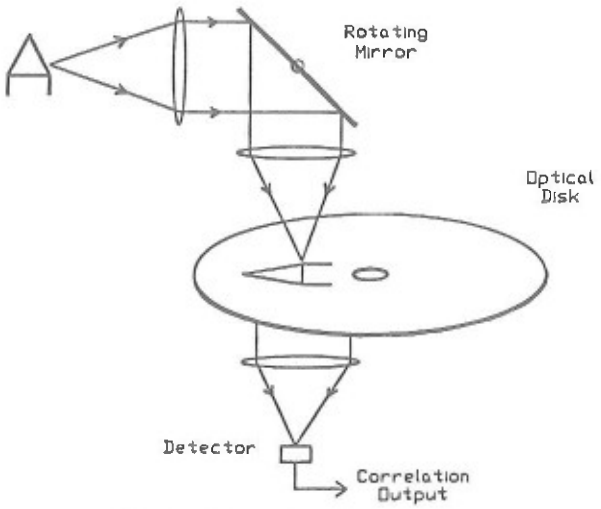
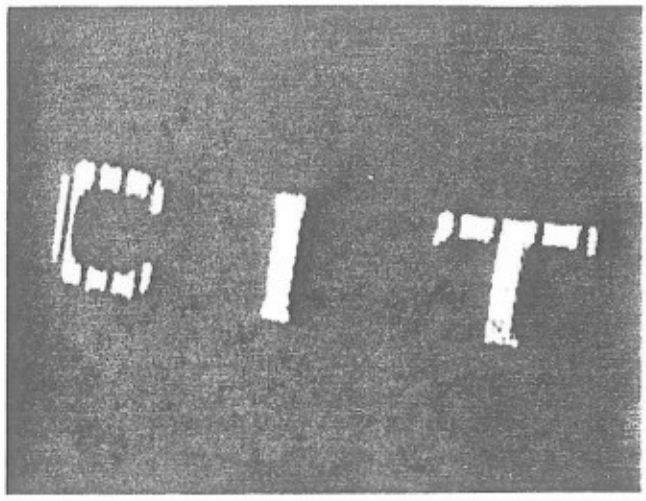
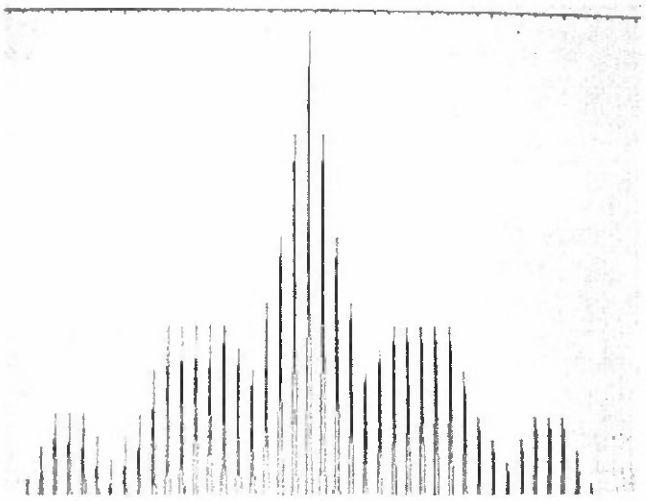


Fig. 25. Rotating mirror correlator.

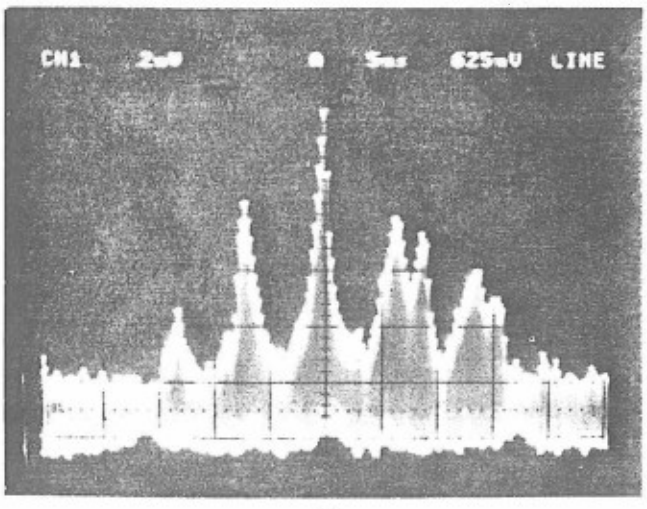
This system is capable of operating with incoherent illumination. One critical limitation, however, is its speed. The speed in this architecture is dictated primarily by the rates of the relative scanning mechanisms. Specifically, to generate an accurate correlation signal, the radial scan time must be less than the time it takes the references to rotate by 1 pixel. The correlation rate is thus limited by factors such as reference image, pixel size, radial scan rate, disk rotation speed, and ultimately by illumination level and disk efficiency. For our experiment, the speed of the rotating mirror is the limiting factor, and it results in a corelation rate of 400 correlations/s for 100 × 100 pixel images. This system, although significantly slower than FT based systems, provides a simple solution to the FT correlator alignment problems by operating in the image plane. As with any incoherent correlator, the present system is a unipolar architecture and some bias removal mechanism is necessary to retrieve a bipolar or high SNR correlation signal. These bias removal techniques have been discussed at length in the literature.[24]



(a)



(b)



(c)

Fig. 26. Rotating mirror correlator results: (a) reference image recorded on Sony disk; (b) computer-generated autocorrelation signal; and (c) optical system output.

D. Acoustooptic Correlator

The most obvious way to improve the speed of the above system is to increase the speed of the radial scanning mechanism. Since the speed of commercially available polygon mirror based scanners is limited to about 40 kHz, we have considered instead the system of Fig. 27 which utilizes an acoustooptic (AO) device as the radial scanner. The AO scanner can achieve scan rates up to 10 MHz. In this architecture, a chirp signal propagating in the AO device generates a moving cylindrical lens with power in the horizontal dimension. This moving cylindrical lens becomes part of the system that images the input onto the disk. Consequently, as the AO lens moves horizontally, the image formed on the disk is scanned radially. The orthogonal scanning is achieved by disk rotation as before and the light collected by the detector once again represents the desired correlation signal.

The correlation rate in this system is still constrained by the radial scan speed; however, since this scanning is generated by virtue of the propagation of a RF chirp in the Bragg cell, the resulting correlation rate is much higher than before. The RF chirp parameters are chosen so as to utilize as much AO space–bandwidth product as the input image requires, while minimizing scan time. Specifically,

$$t_C = t_{AO}(SBP_{IN}/SBP_{AO}), \tag{33}$$

where t_C is the required RF chirp duration, t_{AO} is the AO aperture, and SBP_{IN} and SBP_{AO} are the input and AO space–bandwidth products, respectively. We have built this system using a TeO_2 AO cell with a 70 μs aperture and a RF chirp centered at 40 MHz with a chirp rate of $\approx$ 4 MHz/μs. Using the above equation with SBP_{IN} = 100 and SBP_{AO} = 1000 the required chirp duration was calculated to be 7 μs. A SAW device was used to generate the desired chirp signal. The resulting radial scan rate of 1/7 μs $\approx$ 140 kHz, yields a correlation rate of 1400 correlations/s. Again, this correlator is incoherent, but quasimonochromatic light would be required because of the wavelength sensitivity of the AO lens. The impulse response of the AO lens scanner is shown in Fig. 28. The image of an input transparency was formed on a CCD using one cylindrical lens with power in the horizontal dimension and the AO lens for vertical imaging. The input illumination was pulsed so that the AO lens might be frozen in various vertical positions. The delay between the onset of the RF chirp and the laser diode pulse determines the position of the image on the CCD. As can be seen from the figure, the AO lens imaging characteristics are quite good. The output of the AO lens correlator is shown in Fig. 29. The input to this system was once again the transparency of the acronym CIT and the reference was a duplicate CIT written on an Optotech WORM disk using a simple recording system which we built. The correlator output is, therefore, the 2-D autocorrelation of the input image. The reference image shown in Fig. 29(a) is relatively large so that a radial scanning distance of 2 cm was

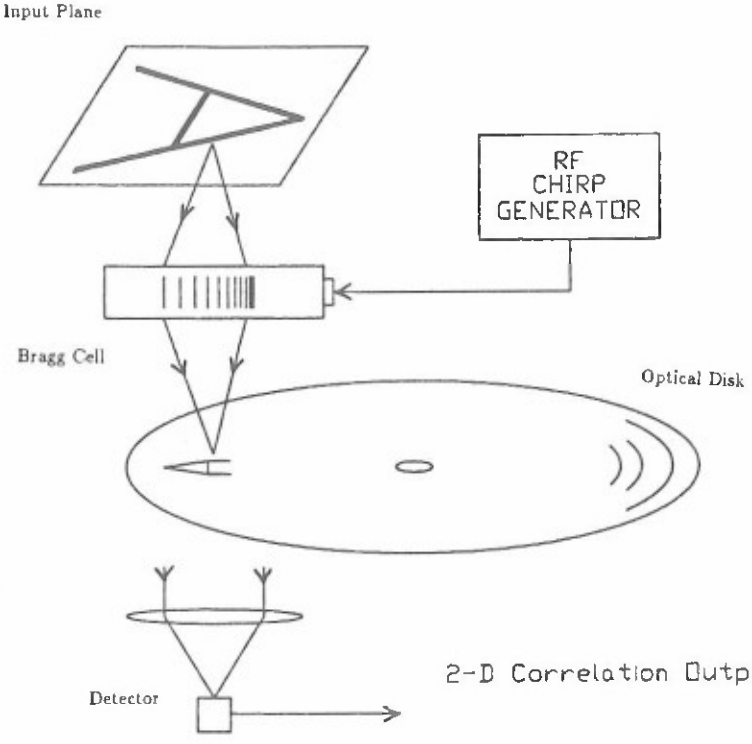


Fig. 27. Moving AO lens correlator.

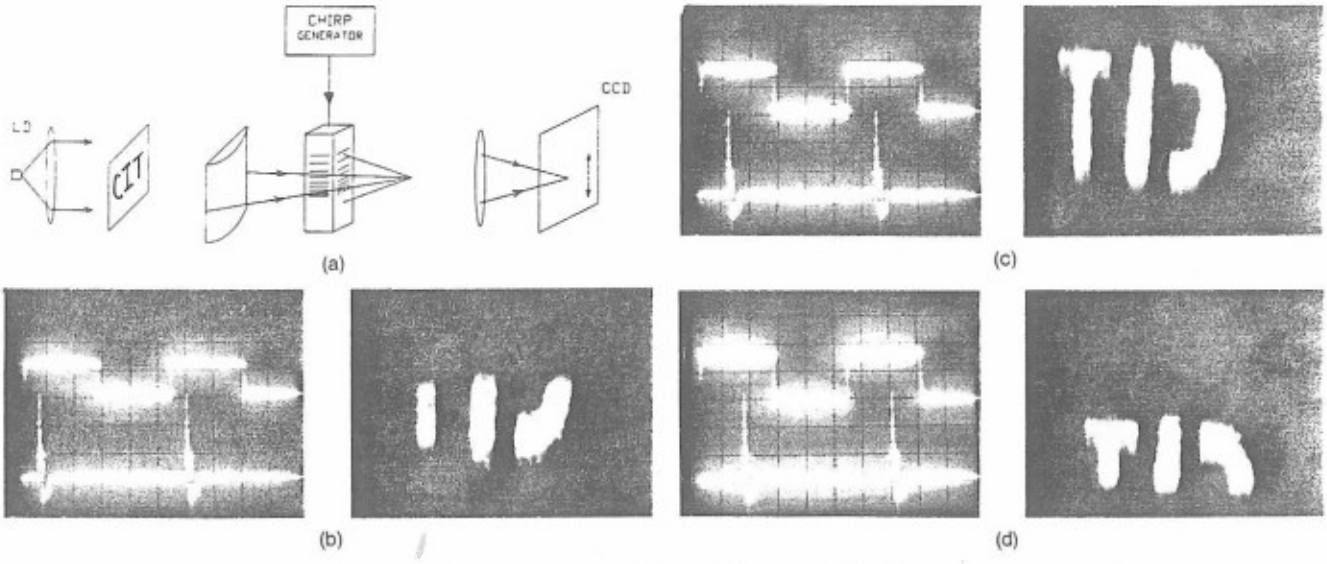
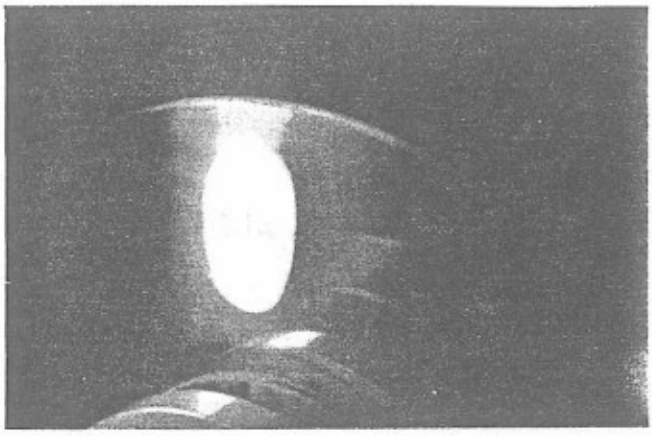


Fig. 28. Impulse response of the AO lens scanner: (a) optical system used to measure impulse response; (b)–(d) image formed on CCD for various delay times Δ, where Δ is the time between the leftmost edge of the chirp gate (upper trace) and the laser diode trigger (lower trace), (b) $\Delta = 2.5\ \mu s$, (c) $\Delta = 5.0\ \mu s$, (d) $\Delta = 8.0\ \mu s$.

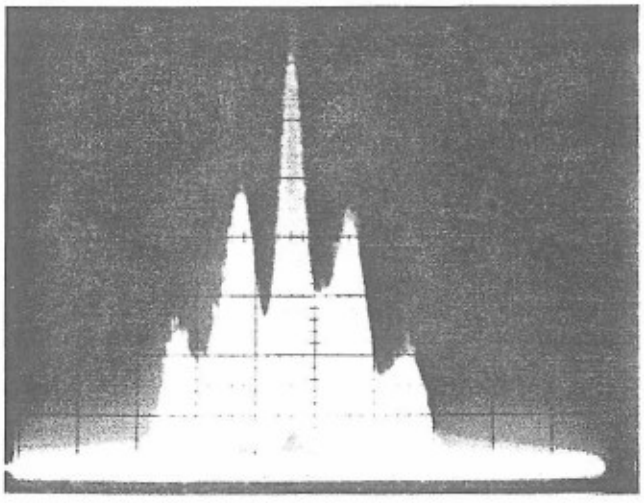
required to generate an accurate correlation signal. As can be seen from Figs. 29(b) and 29(c), the optical system output agrees well with the predicted autocorrelation signal of Figure 26(b).

VI. Conclusions

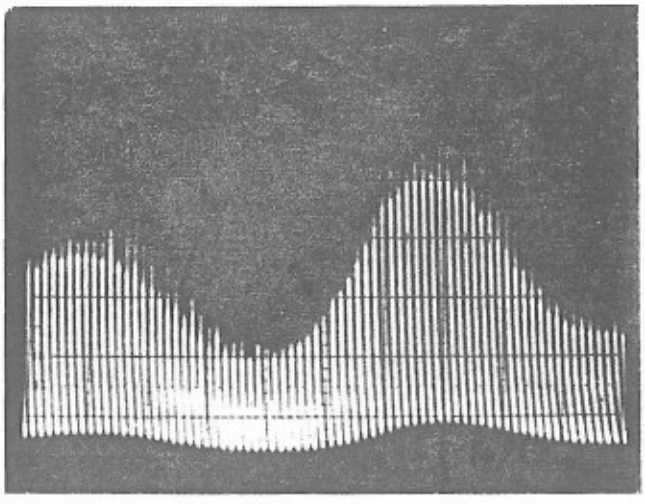
This paper describes the use of optical memory disks in optical computing and optical information processing systems. The large SBP ($\approx 10^{10}$), simple computer addressability, natural scanning mechanism, and parallel accessibility are all features making the optical disk a candidate for use as both memory and SLM in these systems. The sampled format recording scheme results in across track coherence, facilitating the storage of 2-D data on the disk. This format also results in the absence of grooves on the disk, which eliminates



(a)



(b)



(c)

Fig. 29. Output of AO lens correlator: (a) reference on disk; (b) optical system output; and (c) magnified version of (b) to resolve individual radial scan peaks.

sampling of the bias reflectivity thereby increasing image plane contrast. The third attractive characteristic of the disks we used is the optical flatness of the glass coating material. This allows these disks to be used in coherent processing such as holographic reconstruction and complex spatial filtering.

Parallel optical access to images and holograms stored on disks provides the possibility of implementing specialized parallel computing schemes such as database machines, image correlators, and optical disk/VLSI hybrid neural networks. We have demonstrated several of these systems and have shown the potential advantage of such systems over their electronic counterparts. Owing to the maturity of optical disk technology, the architectures described here are feasible using existing disk systems and readily available supporting devices.

This work is supported in part by a grant from the Army Research Office. The authors would like to thank Sony for the disk system used in this work and for the generous support of Seiji Kobayashi during his stay at California Institute of Technology. Special thanks to Adolf Lohmann for many helpful discussions regarding optical disk based computer generated holography. Thanks to Charlie Stirk for his assistance in the design and testing of our VLSI neural net chip. Alan Yamamura is supported by a fellowship from the Fanny and John Hertz Foundation.

Appendix: Diffraction Efficiency for Schlieren Imaging

We calculate, as an example, the diffraction efficiency of the optical disk in a schlieren imaging system. We start by modeling the disk surface according to the following equation:

$$i(x,y) = r_0 + (r_1 - r_0)\left[b(x,y)\sum_{n,m}\delta(x - n\Delta_x, y - m\Delta_y)\right] \otimes \mathrm{circ}\left(\frac{r}{\Delta_r}\right)$$

$$+ (r_2 - 2r_1 + r_0)\sum_{n,m}\left\{[b(x,y)\delta(x - n\Delta_x, y - m\Delta_y)] \otimes \mathrm{circ}\left(\frac{r}{\Delta_r}\right)\right\}$$

$$\times\left\{[b(x,y)\delta(x - (n+1)\Delta_x, y - m\Delta_y)] \otimes \mathrm{circ}\left(\frac{r}{\Delta_r}\right)\right\}, \quad \text{(A1)}$$

where $r^2 = x^2 + y^2$, $b(x,y)$ is the desired binary image, Δ_x the along and Δ_y the across track spacing, Δ_r the radius of the written spots, and r_i the complex amplitude reflectivity of areas written i times. Because spots written by the Sony system have a constant angular separation along track, Δ_x is actually a function of radial position on the disk and varies between Δ_r and $2\Delta_r$. The third term in Eq. (A1) represents the overlap of adjacent spots in the along track direction when $\Delta_x < 2\Delta_r$ (Fig. 30). We otherwise ignore track curvature for now and assume that the pixels lie on a Cartesian grid. The effects of track curvature are analyzed in Secs. II and III.

Because the Sony write once material nearly saturates after a single exposure to the write beam such that $r_2 \approx r_1$, we drop the third term of Eq. (A1) and

account instead for the overlap by modifying the spot shape in the second term by assigning half of the overlap region to each neighboring spot, as in Fig. 30. In this case, the mathematical expression for the spot becomes [circ × rect] and Eq. (A1) reduces to the following:

$$i(x,y) = r_0 + (r_1 - r_0)\left[b(x,y)\sum_{n,m}\delta(x - n\Delta_x, y - m\Delta_y)\right] \otimes \left[\mathrm{circ}\left(\frac{r}{\Delta_r}\right)\mathrm{rect}\left(\frac{x}{\Delta_x}\right)\right]. \quad \text{(A2)}$$

Strictly speaking, the above equation models the reflectivity incorrectly wherever a written spot is adjacent along track to an unwritten one. Assuming, however, that along track spatial frequencies in the image are low compared to the sampling frequency Δ_x^{-1}, we expect separate clusters of written and unwritten spots, in which case there would be little energy in a term accounting for written spots next to unwritten ones.

Using Eq. (A2) to model the reflectivity of the disk surface, we find the following Fraunhofer diffraction pattern:

$$I(u,v) = r_0\delta(u,v) + (r_1 - r_0)\left[B(u,v)\otimes\frac{1}{\Delta_x\Delta_y}\sum_{n,m}\delta\left(u - \frac{n}{\Delta_x}, v - \frac{m}{\Delta_y}\right)\right] \times\left[\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right]_{\rho^2=u^2+v^2} \quad \text{(A3)}$$

As expected, the sampling in the image plane corresponds to convolving the image spectrum with an array of impulses resulting in an array of image spectra or diffraction of the image into multiple orders. Schlieren imaging achieves high contrast because the pixels sample only the image and not the background, represented by r_0, sending all the energy in the bias to the zero-order.

The fraction of incident light that goes into the n,mth diffraction order is given by integrating the magnitude squared of the appropriate term from Eq. (A3) as follows:

$$H_{nm} = \left|\frac{\Delta_r(r_1 - r_0)}{\Delta_x\Delta_y}\right|^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|B\left(u - \frac{n}{\Delta_x}, v - \frac{m}{\Delta_y}\right) \times\left[\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right]\right|^2 du\,dv. \quad \text{(A4)}$$

The shape of the written spots determines the characteristic $[J_1 \otimes (\mathrm{sinc} \times \delta)]$ envelope which modulates the entire diffraction pattern in Eq. (A3). Assuming that the pixels sufficiently oversample the input image, the envelope is nearly constant over the image spectrum allowing us to simplify Eq. (A4):

$$H_{nm} \approx \left|\frac{(r_1 - r_0)}{\Delta_x\Delta_y}\right|^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\left(u - \frac{n}{\Delta_x}, v - \frac{m}{\Delta_y}\right)\right|^2 du\,dv \times\left|\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n/\Delta_x, v=m/\Delta_y}. \quad \text{(A5)}$$

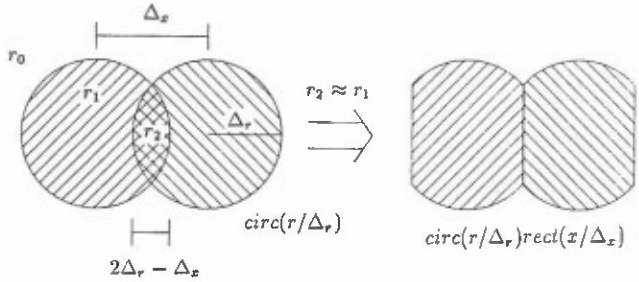


Fig. 30. Model of spot shapes.

The remaining integral corresponds to the energy in the spectrum of the image which is equivalent to the energy in the image itself. Thus, the fraction of incident light energy that goes into the n,mth diffraction order is given as follows:

$$H_{nm} \approx \left|\frac{r_1 - r_0}{\Delta_x\Delta_y}\right|^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|b(x,y)|^2 dx\,dy \times\left|\frac{\Delta_r J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n/\Delta_x, v=m/\Delta_y}. \quad \text{(A6)}$$

Defining useful light as the total energy going into image spectra of all orders, we estimate the light captured by imaging a single order as a fraction of useful light:

$$\frac{H_{nm}}{\sum_{n',m'=-\infty}^{\infty}H_{n'm'}} = \frac{\left|\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n/\Delta_x, v=m/\Delta_y}}{\sum_{n',m'=-\infty}^{\infty}\left|\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n'/\Delta_x, v=m'/\Delta_y}}. \quad \text{(A7)}$$

Since the denominator is the sum of the squares of the Fourier series coefficients of an image with every spot written with unity amplitude, it is equivalent to the fraction of the disk covered by spots if all were written. This fraction varies between 0.877 at the innermost radius and 0.785 at the outermost. We can now simplify Eq. (A7) as follows:

$$\frac{H_{nm}}{\sum_{n',m'=-\infty}^{\infty}H_{n'm'}} = \frac{\left|\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n/\Delta_x, v=m/\Delta_y}}{\frac{1}{\Delta_x\Delta_y}\int_{-\Delta_x/2}^{\Delta_x/2}\int_{-\Delta_y/2}^{\Delta_y/2}\left|\mathrm{circ}\left(\frac{r}{\Delta_r}\right)\mathrm{rect}\left(\frac{x}{\Delta_x}\right)\right|^2 dx\,dy}, \quad \text{(A8)}$$

$$= \frac{\left|\Delta_r\frac{J_1(2\pi\Delta_r\rho)}{\rho}\otimes\Delta_x\,\mathrm{sinc}(\Delta_x u)\delta(v)\right|^2_{u=n/\Delta_x, v=m/\Delta_y}}{\frac{1}{\Delta_x\Delta_y}\left[\frac{2\Delta_r^2}{\Delta_x\Delta_y}\sin^{-1}\left(\frac{\Delta_x}{2\Delta_r}\right) + \frac{\Delta_r}{\Delta_y}\sqrt{1 - \frac{\Delta_x}{2\Delta_r}}\right]}. \quad \text{(A9)}$$

Thus, Eq. (A6) shows the amount of light as a fraction of incident light that goes into the image spectrum

in each order. Equation (A9) shows the amount of image light in each order as a fraction of image light in all orders. These equations can be used to estimate the diffraction efficiency of the Sony write-once disk in a schlieren imaging system. The disk diffraction efficiency for other applications and/or disks can be estimated in a similar fashion given a model for the reflection or transmission pattern of the disk and parameters for the appropriate disk characteristics.

References

1. R. Bartolini, H. Weakliem, and B. Williams, "Review and Analysis of Optical Recording Media," Opt. Eng. 15, 99–108 (1976).
2. Y. Abu-Mostafa and D. Psaltis, "Optical Neural Computers," Sci. Am. 255, 88–95 (1987).
3. D. Psaltis, A. A. Yamamura, M. A. Neifeld, and S. Kobayashi, "Parallel Readout of Optical Disks," in *Technical Digest, Topical Meeting on Optical Computing* (Optical Society of America, Washington, DC, 1989), pp. 58–61.
4. L. Giles and B. K. Jenkins, "Models of Parallel Computation and Optical Computing," in *Technical Digest, OSA Annual Meeting* (Optical Society of America, Washington, DC, 1986), paper ML1.
5. Y. Nakane *et al.*, "Principle of Laser Recording Mechanism by Forming an Alloy in the Multilayer of Thin Metallic Films," Proc. Soc. Photo-Opt. Instrum. Eng. 529, 76–81 (1985).
6. D. Psaltis, E. G. Paek, and S. S. Venkatesh, "Optical Image Correlation with a Binary Spatial Light Modulator," Opt. Eng. 23, 698–704 (1984).
7. J. H. Rilum and A. R. Tanguay, Jr., "Utilization of Optical Memory Disks for Optical Information Processing," in *Technical Digest, OSA Annual Meeting* (Optical Society of America, Washington, DC, 1988), paper M15.
8. J. F. Jarvis, C. N. Judice, and W. H. Ninke, "A Survey of Techniques for the Display of Continuous Tone Pictures on Bilevel Displays," Comput. Graphics Image Process. 5, 13–40 (1976).
9. Y. Tsunoda, K. Tatsuno, K. Kataoka, and Y. Takeda, "Holographic Video Disk: An Alternative Approach to Optical Video Disks," Appl. Opt. 15, 1398–1403 (1976).
10. I. Satoh and M. Kato, "Holographic Disk Recording of Digital Data with Fringe Stabilization," Appl. Opt. 27, 2987–2992 (1988).
11. T. Yatagai, J. G. Camacho-Basilio, and H. Onda, "Recording of Computer-Generated Holograms on an Optical Disk Master," Proc. Soc. Photo-Opt. Instrum. Eng. 1052, 119–124 (1989).
12. T. Inagaki, "Hologram Lenses Lead to Compact Scanners," IEEE Spectrum 26, 39–43 (1989).
13. B. Brown and A. Lohmann, "Complex Spatial Filtering with Binary Masks," Appl. Opt. 5, 967–969 (1966).
14. W.-H. Lee, "Binary Computer Generated Holograms," Appl. Opt. 18, 3661–3669 (1979).
15. G. Tricoles, "Computer Generated Holograms: An Historical Review," Appl. Opt. 26, 4351–4360 (1987).
16. A. Lohmann, U. Erlangen-Nuremburg; personal communication.
17. J. Alspector and R. B. Allen, "A Neuromorphic VLSI Learning System," *Advanced Research in VLSI Processes 1987 Stanford Research Conference* (MIT Press, Cambridge, 1987), pp. 313–349.
18. Special Issue on Neural Networks, *Applied Optics* 26, (1 Dec. 1987).
19. J. H. Kim, S. H. Lin, J. Katz, and D. Psaltis, "Monolithically Integrated 2-D Arrays of Optoelectronic Devices for Neural Network Applications," Proc. Soc. Photo-Opt. Instrum. Eng. 1043, 44–52 (1989).
20. A. VanderLugt, "Signal Detection by Complex Spatial Filtering," IEEE Trans. Inf. Theory IT-10, 139 (1964).
21. D. Psaltis, M. A. Neifeld, and A. A. Yamamura, "Optical Disk Based Correlation Architectures," in *Technical Digest, Topical Meeting on Optical Computing* (Optical Society of America, Washington, DC, 1989), pp. 206–209.
22. D. Psaltis, M. A. Neifeld, and A. A. Yamamura, "Image Correlators Using Optical Memory Disks," Opt. Lett. 14, 429–431 (1989).
23. J. Yu, "Optical Processing Using Photorefractive Crystals," Ph.D. Thesis, California Institute of Technology (1988), Chap. 5.
24. D. Psaltis, "Incoherent Electrooptic Image Correlator," Opt. Eng. 23, 12–15 (1984).
25. A. D. Mikaelyan, A. Vanin, E. D. Gulanyan, and S. Prokopenko, "Holographic Disk for Data Storage," *Sov. J. Quantum Electronics*. 170(5), 680–687 (1987).

Optical Computing and Processing, 1991, volume 1, number 1, pages 3–12

Optical-disk based artificial neural systems

ALAN A. YAMAMURA, MARK A. NEIFELD, SEIJI KOBAYASHI AND DEMETRI PSALTIS

Optical disks provide a mature technology for the storage and implementation of the connection patterns required in artificial neural systems. In this paper, we briefly characterize optical disks before describing and presenting experimental results from two optical-disk based neural networks: a character recognition system and a multilayer feedforward neural network simulator.

1. Introduction

The majority of today's artificial neural network models [1] consist of two elements: a large number of simple processing elements (*neurons*) and the connections (*synapses*) between them. The pattern of connections determines the functionality of the network and is often generated through learning and modified through adaptation. Both electronic and optical technologies are currently being used to implement artificial neural networks. The key advantages of optics are its abilities to provide the large number of connections required by many neural network models and efficiently store the connection pattern specifications. While electrical signals must travel on physical wires that consume space, optical signals can propagate through free space, and optical memories such as disks are valued for their high density and high capacity data storage. Optical disks can also act as parallel readout storage elements and spatial light modulators [2]. Recently, optical disks have been used as a component in the implementation of neural networks [2–4]. In this paper, we describe two optical-disk based artificial neural systems. The first is an optical character recognition system that uses the disk both to store and to implement connections optically. The second is a multilayer feedforward neural network processor that uses the optical readout capability of the disk to transfer the connection patterns in parallel from the disk to an optoelectronic processing chip.

Received 5 June 1990.

Authors' address: California Institute of Technology 116–81, Pasadena, CA 91125, U.S.A.

Figure 1. Greyscale image recorded on an optical disk by area modulation.

2. The optical disk

Using a prototype Sony system, we have recorded information in the form of images and holograms on reflective disks, 12 cm in diameter [2]. Data is recorded with 1 μm resolution as variations in the surface reflectivity of the disk yielding a storage capacity or space–bandwidth product (SBP) of well over 10^9 bits (that is, thousands of images or holograms of a million pixels each) per disk side.

Figure 1 shows an image stored as a 2-D pixel array on the disk surface. Although the individual pixels can

represent only binary information, greyscales are recorded in this image with area modulation techniques. By imaging the light reflected by an area of the disk under uniform illumination, we can read out a 2-D array of data in parallel. In this case, the disk simply acts as an optical storage medium with parallel readout. The disk can also be used to implement weighted connections using a vector–matrix multiplication architecture (figure 2(*a*)). Here, a vertical array of N neurons can be fully connected to a horizontal array of M neurons using a disk and a pair of anamorphic imaging lenses. The connection pattern is specified by an array of $N \times M$ superpixels recorded on the disk. Light emitted by each neuron in the vertical array is imaged by the first lens onto a row of the connection matrix; light reflected by each column of the matrix is integrated by the second lens onto a neuron in the horizontal array.

Information may also be stored in the form of computer generated holograms written on the disk. As in the case of the images, holograms on the disk can be used either to store data for parallel optical readout or to implement weighted connections. For parallel optical readout, we simply encode the hologram so that it reconstructs the desired pattern of points under illumination by a planewave reference. For weighted connections, the hologram is encoded to partition light emanating from each neuron and distribute it to the other neurons in a programmable fashion (figure 2(*b*)).

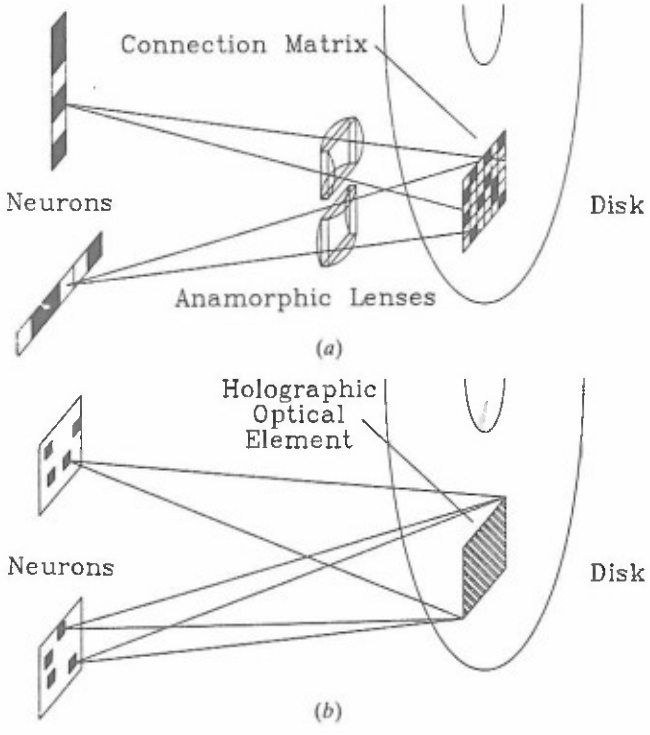


Figure 2. Optical implementation of neural connections: *(a)* vector–matrix and *(b)* holographic.

Holographic storage has a number of potential advantages. First, optical elements may be encoded into the holograms to allow lensless readout or interconnection. Second, holograms represent information in a distributed fashion, resulting in slow degradation of data or connections as the number of defects in the hologram increases. Third, shifting a Fourier transform hologram (FTH) results only in phase changes in the reconstruction; thus for small rotations the light reflected by a FTH remains aligned with a detector array as the disk rotates [5].

Unfortunately, holographic recording also has a number of disadvantages. A large amount of SBP must be devoted to record and readout a hologram accurately. Superpixels (typically groups of 4–64 or more pixels) are usually required to record the amplitude and the phase of each sample point in the hologram. Additional SBP must be sacrificed to incorporate spatial carriers or holographic optical elements, if desired, in the holograms, as well as to improve the signal-to-noise ratio (SNR). Figure 3 shows the desired object and experimental reconstruction from computer generated holograms recorded on the disk as we vary the number of pixels in the reconstruction. Starting from an initial value of 32 'on' pixels, the SNR decreases steadily. With 512 'on' pixels, the signal is almost completely lost in the noise. Over a million pixels on the disk were used to encode the 1024 pixels in the reconstruction, corresponding to an additional factor of 1024 in required SBP for holographic recording when compared with imaging.

3. Optical character recognition system

The first system we describe is an optical disk based system for the recognition of handwritten numerals. The recognition scheme is based on a K nearest neighbour strategy that uses a template library of 650 exemplars. The optical system compares an unknown input against the template library at a demonstrated rate of 26 000 comparisons per second. Although shift, rotation, and scale invariances may be effectively eliminated through normalization procedures [6], author-dependent distortions are often dealt with by statistical techniques based on a large training set of exemplar patterns. Such approaches are typically based on computationally intensive algorithms requiring a great deal of both time and memory [7]. Whereas such computationally intensive algorithms are difficult to realize on conventional computers, parallel access optical storage technologies such as the optical disk, can provide an efficient implementation. Optical parallel access to information stored on the disk provides a mechanism for high data retrieval rates and concomitantly large

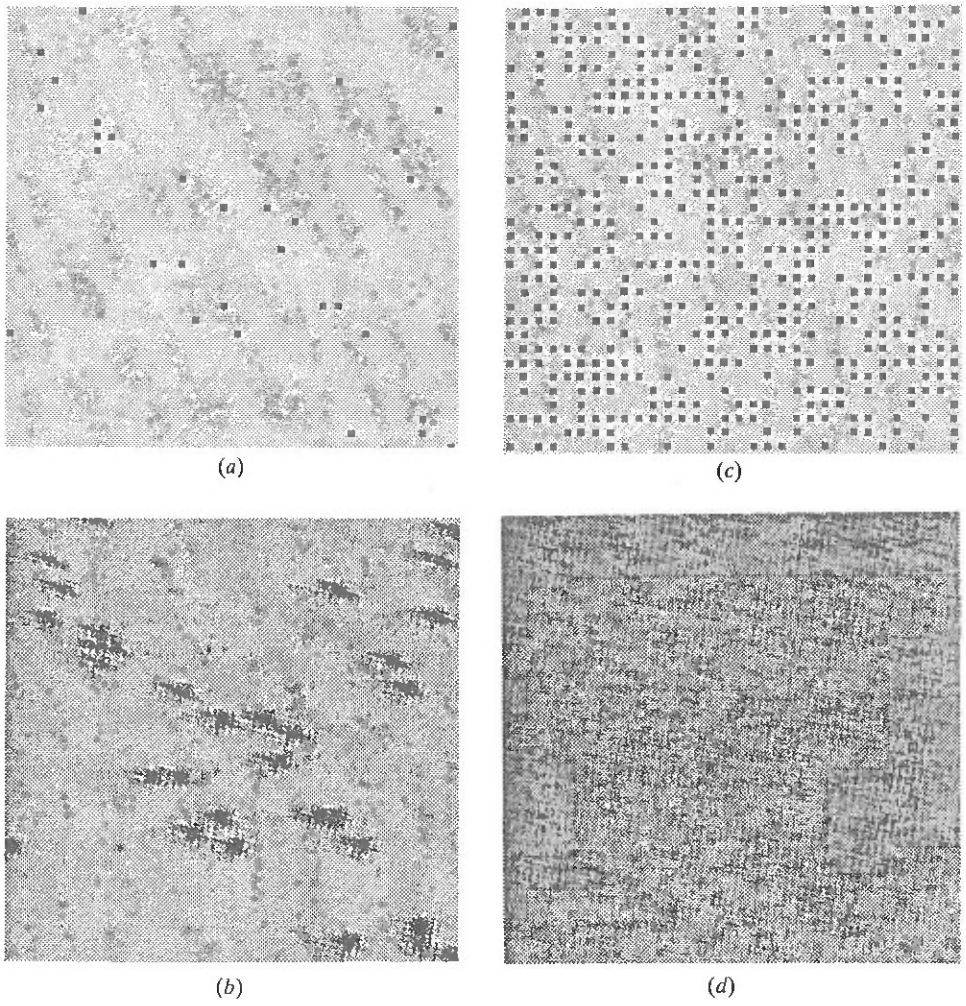


Figure 3. Desired *(a, c)* objects and *(b, d)* experimental reconstructions of CGHs on disks with *(a, b)* 32 and *(c, d)* 512 pixels, respectively.

processing speeds [2]. The combination of parallel access and large storage capacity makes optical-disk based architectures well suited to the efficient implementation of several of the traditional pattern recognition paradigms as well as neural network models. In this paper we will demonstrate experimentally an optical system which performs handwritten numeral recognition using the K nearest neighbour algorithm (KNN).

Given a set of M training vectors which we will refer to as templates, the KNN algorithm classifies an unseen vector according to the class with the greatest representation among its K nearest neighbours in the template set [8]. To realize this algorithm we must compute the M distances $|x-x^i|$, for $i=1,\ldots,M$ where x is the unknown input vector and the x^is are the stored templates. The success of this algorithm depends on how well the template set represents the underlying problem at hand. As M becomes very large, the probability of error for the KNN algorithm is known to approach the optimum value, but the computational requirements can become impractical for conventional computers. The optical system described here, however, is capable of performing ≈ 40 million such comparisons per second, and it is possible to store $\approx 10^6$ templates of 10^4 bits each on a single disk. Thus, this optical technique may be able to push the boundaries of the practicality of the KNN rule well beyond what is currently possible.

A database of 950 16×16 handwritten numerals (95 per class) was used to construct the training and testing sets for our experiments: 65 vectors were chosen randomly from each of ten classes to generate the 650 element training set (figure 4); the remaining 300 vectors were used as a testing set. It was found that for a fixed template set size of 650 vectors, the recognition performance of the KNN algorithm was improved dramatically if the 16×16 binary input vectors were first normalized for position and scale. Accordingly, a preprocessing step consisting of a centring operation followed by a scaling of the centred 16×16 character to a 10×10 window was performed. The 10×10 templates were unraveled and stored as 100 dimensional vectors on the optical disk. Along with each binary template x, its complement $\bar{x}$ (that is $\bar{x}+x=(1,1,\ldots,1)$) was

generated and stored on the disk. Thus, we can implement bipolar valued templates by subtracting the inner products between the same input and the two stored vectors. The 650 templates and their complements were recorded as 1300 radial lines on an optical disk, which serves as a parallel access template library in our experiment. The remaining 300 characters were preprocessed in the same way and recorded on transparencies to serve as testing inputs to our optical system.

The optical system is shown in figure 5. An image of the input vector located on an input spatial light modulator (SLM) is formed on the optical disk along a radial line as shown. The light diffracted from the disk is collected on the photodetector and represents the inner product between the input vector and the vector recorded along the illuminated line. As the disk rotates the input is compared against all of the stored templates and an electrical signal representing the result of these comparisons is analyzed by postprocessing electronics to determine the correct classification of the unknown vector. In our experiment, the postprocessing electronics were responsible for sampling the output of the photodetector and using the inner product data to calculate the KNN of the input vector. The inner product based distance metric used is given by:

$$y = \frac{x \cdot (x^i - \bar{x}^i)}{|x^i|^2}, \tag{1}$$

where the $|x^i|^2$s were computed using the optical system output when presented with the input vector having all its components equal to one. The disk rotation rate in our system was 20 Hz; therefore, the number of binary inner products being computed per second was 26 000. The resolution of the optical disk will allow storage of up to 10^6 templates of dimension 100×100. At a 20 Hz rotation rate, this corresponds to a computing rate of 2×10^{11} binary operations per second.

When tested on the 300 remaining vectors, the optical system achieved a recognition rate of 71% using a $K = 5$ KNN algorithm. It should be noted here that the 300 test vectors had not been seen by the system prior to testing. This performance is considerably below the performance of 83% correct classification predicted by a computer simulation. A model of the optical system predicts a 73% rate. This model included various error sources such as beam nonuniformity,

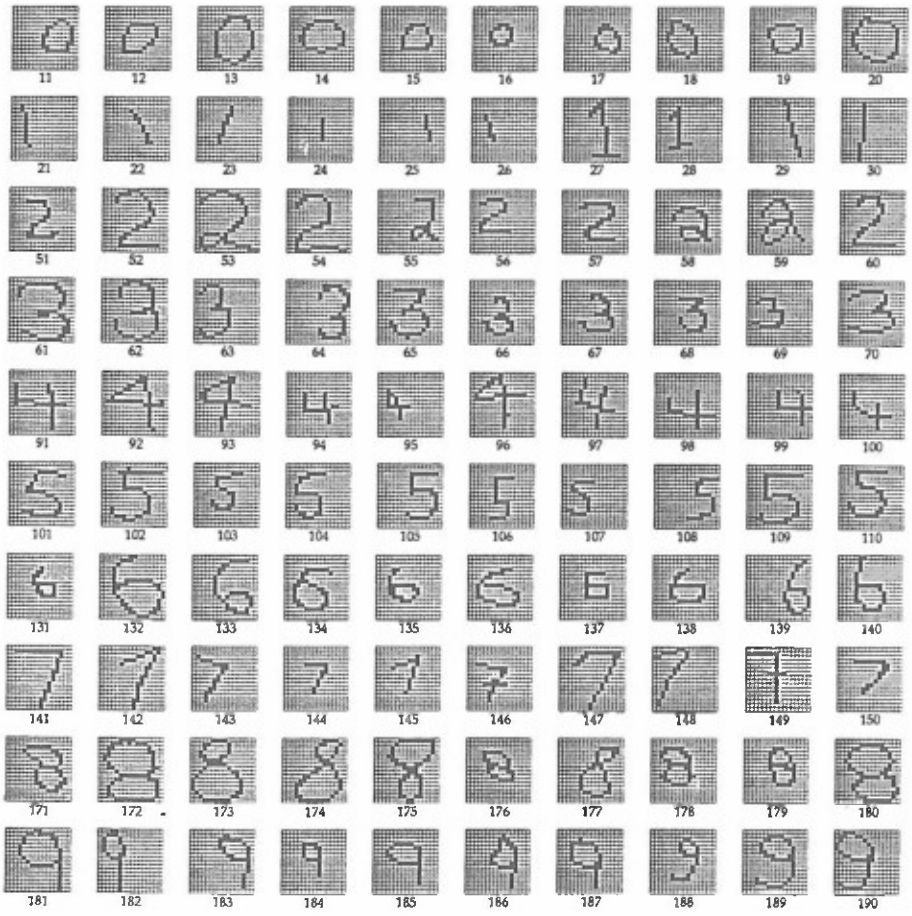


Figure 4. 100 of 540 handwritten character templates used in the character recognition system of figure 5.

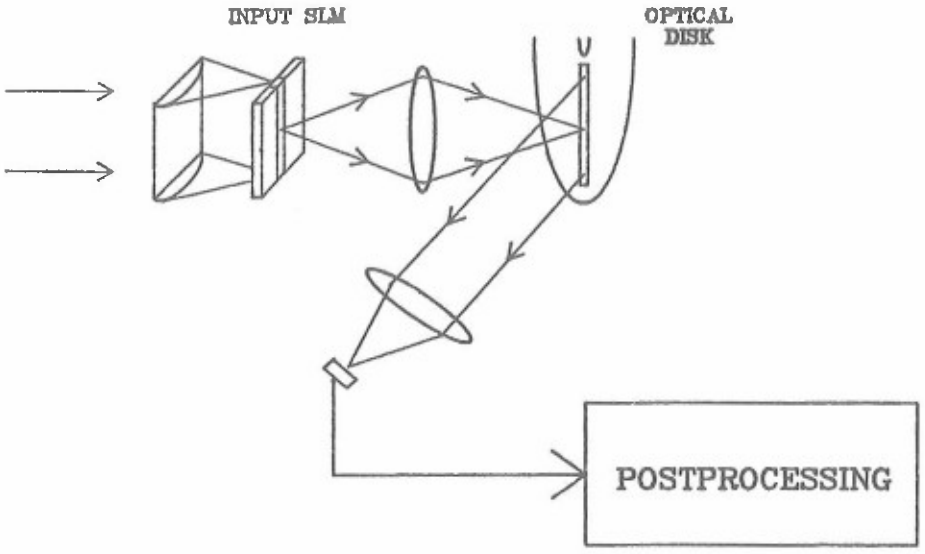


Figure 5. Optical character recognition system.

electrical noise in the postprocessing electronics, sampling phase jitter, and quantization error. The most critical parameter is the input image contrast. These results are summarized in table 1. A plot of recognition rate versus input image contrast is shown in figure 6. The data plotted includes experimental values of the error sources mentioned above in addition to finite contrast. It can be seen from the graph that for our system operating at a measured contrast of $<20:1$, an expected rate of $<75\%$ is obtained in agreement with experiment. It is clear that noise sources only account for $\approx 20\%$ additional recognition error and upon improvement of the input contrast, the optical system is expected to perform near the simulation rate of 83%.

We have demonstrated an optical system capable of comparing an input vector against a large library of stored templates at MHz rates. This system is attractive from the perspective of data reduction in image recognition oriented tasks. The optical system can be envisioned to be an efficient preprocessor for high dimensional input data effectively reducing the dimensionality by projecting the input image onto stored templates or feature vectors stored on the optical disk. The reduced dimensionality output of the optical system is well suited not only to the KNN recognition algorithm demonstrated here, but also to many other pattern recognition and neural network schemes that require the calculation of inner products such as Parzen windows, multilayer networks and associative memories, as well as hypersurface reconstruction networks using radial basis functions [9]. In general, we can envision hybrid pattern recognition systems utilizing high-speed optical preprocessors followed by more conventional electronic computing elements which together will be capable of realizing many different algorithms and networks in a flexible and efficient fashion.

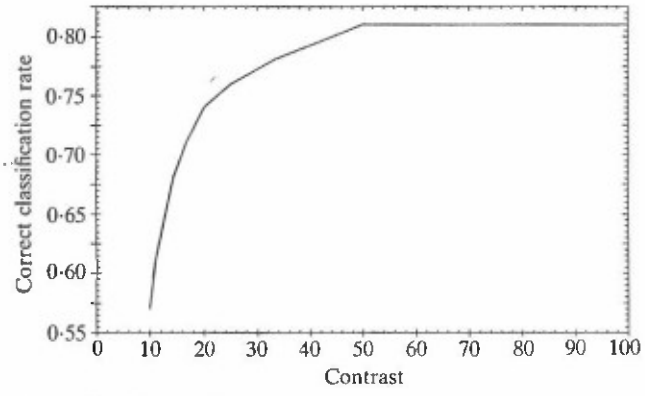


Figure 6. Recognition rate versus image contrast.

Table 1. Comparison of the simulated, experimental and modeled character recognition rates for the digits 0–9.

Class	Simulation	Experiment	Model
0	28	28	28
1	27	19	17
2	24	28	24
3	23	9	15
4	28	24	26
5	24	26	21
6	22	23	20
7	24	26	17
8	23	21	26
9	26	9	24
	83%	71%	73%

4. An optoelectronic multilayer network

A hybrid optoelectronic implementation combines the strength of optics in communications with that of electronics in computation [10]. In this section we present a multilayer feedforward neural network implementation that uses neurons and synapses fabricated on an integrated circuit. Each synapse contains a photodetector thus allowing the weights of the connections to be accessed optically from a disk as shown in figure 7.

Because of its structure, a multilayer feedforward network can be implemented using a chip with only a single layer of neurons and synapses by repeatedly reconfiguring the synapses to implement the succeeding layer of the network before feeding the neuron outputs back to the synaptic inputs. The first advantage of this technique is that a single programmable chip can be used to perform multiple functions. Assuming 10^4 synapses on each chip, a single disk can store almost 10^6 different connection patterns. Depending on the average number of connection patterns required to implement a function, a single chip and disk combination could perform hundreds of thousands of different functions. Almost a million chips with fixed connectivity would be required to provide the same functionality. Secondly, for those tasks that are too large to fit on a single chip or wafer (but can be partitioned into pieces that can be implemented in a serial fashion), reconfigurable connections allow us to implement the entire network on a single chip, thus avoiding potentially long communication delays between separate chips implementing different parts of the network. When reconfigurable synapses are used, a set of weights must be stored for each layer of every function we would like the network to perform. Since the storage of weights on-chip consumes area that could otherwise be used for computation, weight storage on an optical disk can reduce the complexity of the circuit dramatically. Optical storage and readout of the weights shortens the time required to load the weights into the synapses. As the number of neurons N in each layer grows, the number of synapses grows as $O(N^2)$ while the number of pads would at best grow as $O(N)$. Thus $O(N)$ cycles would be required to electronically load the weights for each layer. However, with photodetectors in each synapse, the weights for an entire layer can be read in optically in parallel in one clock cycle via the third dimension.

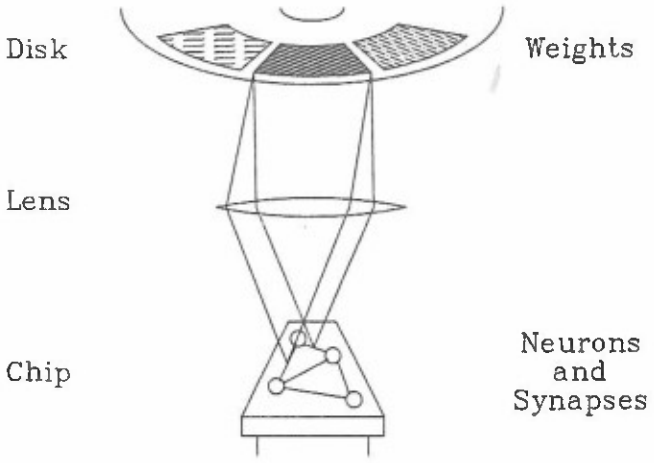


Figure 7. Optical reconfiguration of synaptic weights.

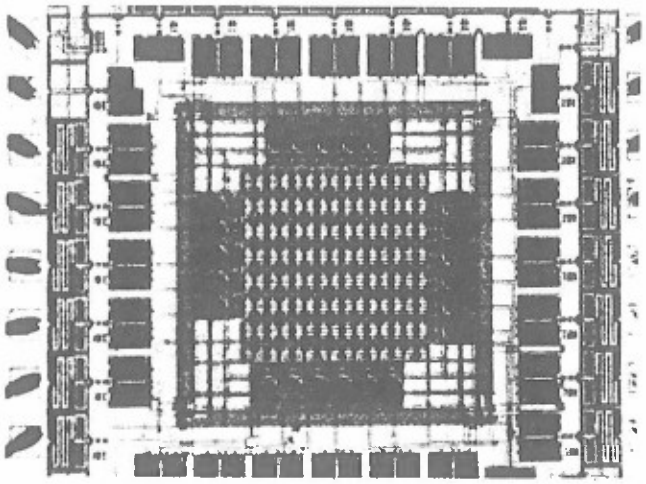


Figure 8. Optoelectronic multilayer feedforward neural network chip.

We have designed and tested an integrated circuit (figure 8) containing two layers of 11 neurons and a 15×15 synaptic array connecting them. (For the purpose of electrical characterization, some synapses do not connect to neurons). Each synapse (figure 9(*a*)) contains a synaptic transistor connecting a pair of neurons, one from each layer. The strength or weight of the connection depends on the gate voltage of this transistor. The gate voltage is determined by a pullup transistor to V_{dd} and a reverse-biased photodiode to ground (G). The weight of a synapse is controlled by adjusting the amount of light striking the photodiode. We operate the synapses in a fashion that implements binary (0, 1) weights. First, we turn on the pullup transistors with a reset signal that precharges the gates of the synaptic transistors thus setting all weights to 1. Next, we selectively illuminate the photodiodes in some of the synapses, discharging the gates of their synaptic transistors and setting the weights of the selected synapses to 0. We could implement analogue weights using the chip designed; however, the storage of analogue weights would consume additional storage space on the optical disk and possibly reduce the data transfer bandwidth.

Figure 9(*b*) shows the neuron circuit. Though this circuit is more complex than the synapse circuit, we can fortunately afford to consume more area with each neuron since there are only N neurons compared with $O(N^2)$ synapses. The voltage input of each neuron is determined by the output voltages of neurons in the previous layer and the strengths of the synaptic connections to those neurons. To generate its output, each neuron applies a hard threshold to the input voltage by comparing it with an adjustable threshold voltage V_{th}; a cross-coupled inverter is used for the voltage comparison. The output of the voltage comparator is V_{dd} if the input voltage is above V_{th} and G if below, thus providing binary outputs.

The switching energy E_s of each synapse is an important parameter which can be used to determine the optical power required to operate the network at a given speed. The switching energy can be determined using the following equation:

$$E_s = IA_p t_s, \tag{2}$$

where I is the intensity of the light uniformly illuminating the chip, $A_p = 474\ \mu m^2$ is the area of photodiode, and t_s is the time required to switch the synapses. Figure 10 shows the current through an array of synapses as a function of the illumination time for a light intensity of $33{\cdot}7\ \mu W\ cm^{-2}$. By averaging measurements for four intensities, we find a switching energy of 0·7 pJ.

The speed at which we can operate the system is limited by the maximum rate at which we can rotate the disk and still maintain alignment between the data from the disk and the detectors on the chip. When imaging data from disk to chip, we can access timing and tracking information stored on the disk through either auxiliary detectors on the chip or the serial read/write head currently used to record information on the disk. The timing information can be used to pulse the readout laser thus 'freezing' the rotation of the disk when the data is in azimuthal alignment with the chip. The tracking information could be used by a servomechanism to move the chip into radial alignment with the data on the disk. A better solution might be to connect the neurons in each layer like a shift-register and electronically 'reposition' the detectors.

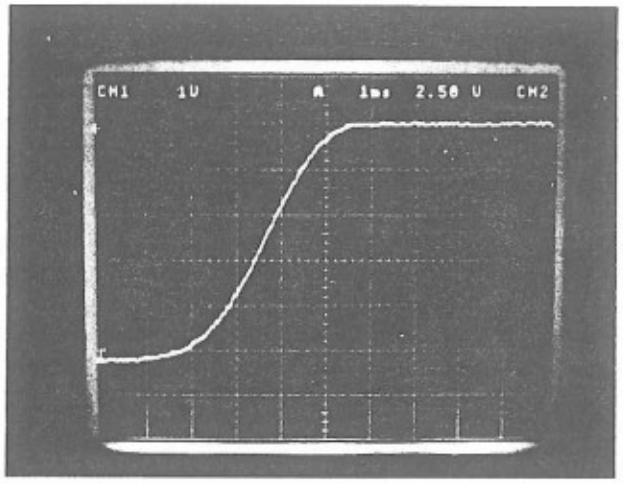


Figure 10. Synaptic current (6·67 μA per vertical division) versus optical illumination time (1 ms per horizontal division).

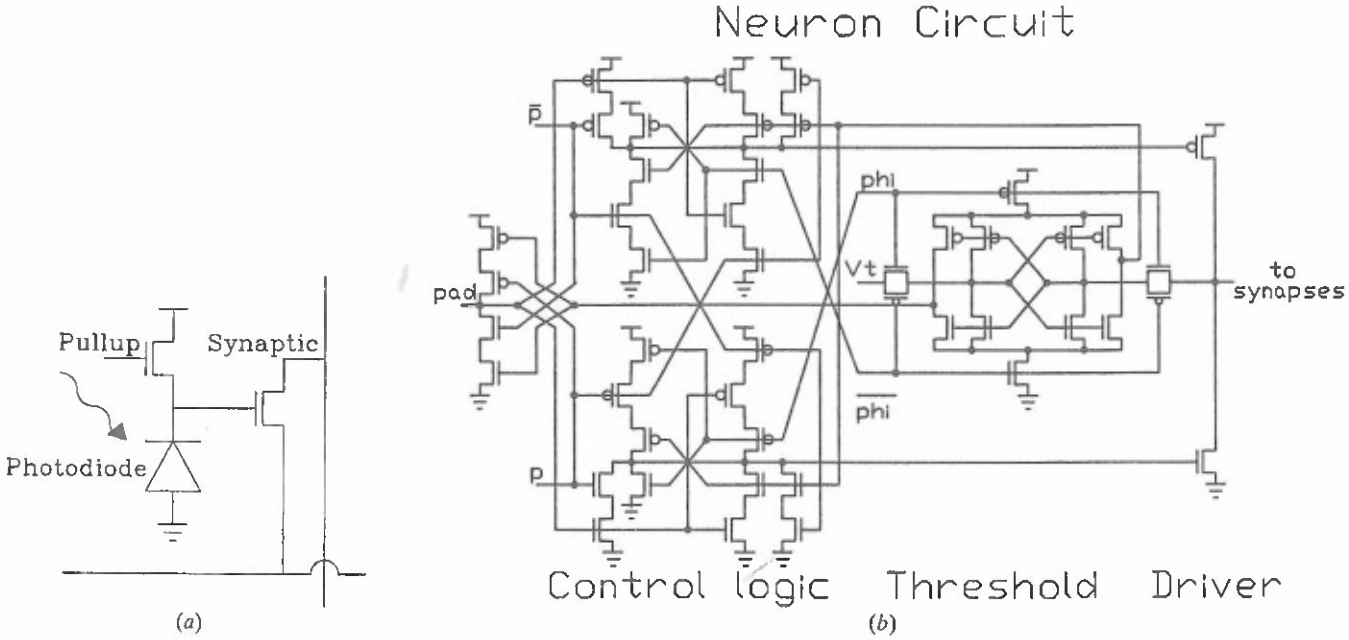
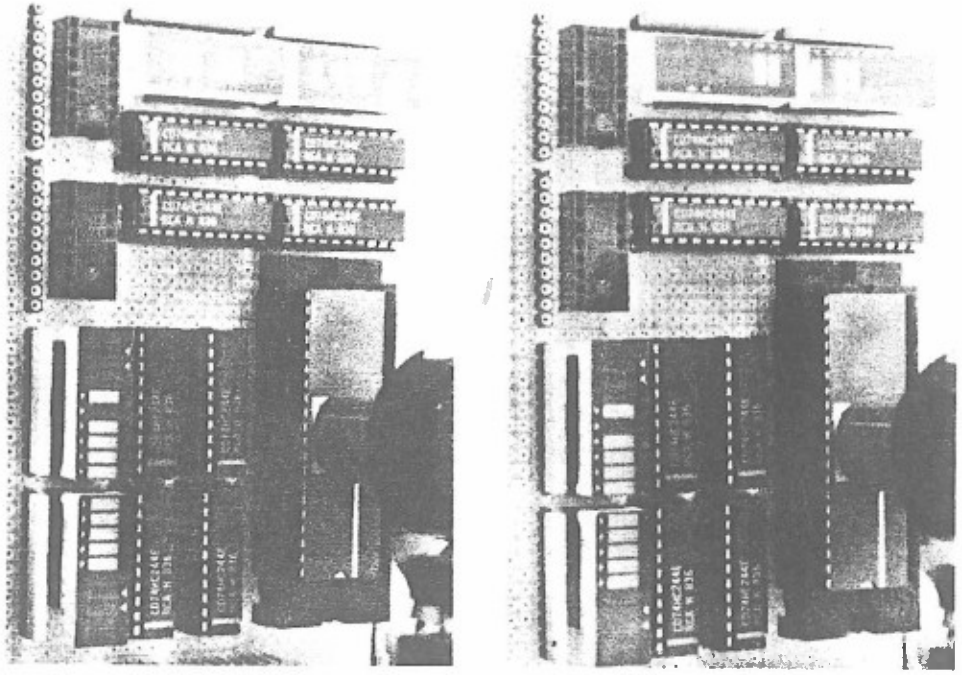


Figure 9. Circuit diagrams: *(a)* synapse, *(b)* neuron.

As mentioned earlier, storage of the weights in the form of a 2-D FTH provides some timing and tracking error tolerance because of the invariance of the intensity pattern of the reconstruction to shifts of the hologram. The major drawback of this scheme is the large cost in SBP required by holographic data storage. It is also possible to compromise between 2-D imaging and holographic readout by recording the data for imaging readout in one dimension and holographic readout in the other by use of an anamorphic optical system. For example, 1-D transforms used in the radial dimension to simplify alignment in the radial direction could be combined with a pulsed laser that provides alignment in the azimuthal direction.

We could potentially spin disks on air bearings at up to 60 000 rpm. If we assume a 100×100 synaptic array on the chip and that we can maintain alignment at this disk rotation rate and use connection patterns stored consecutively on the disk, this rotation rate corresponds to a minimum clock period of 300 ns and a data transfer rate of 35 Gbit s^{-1}. With a neuron switching time of 50 ns and RC delays of 100 ns in the synaptic array, we are left with about 100 ns to switch the synapses. Using a synapse switching energy of 1 pJ, we would then require 100 mW of optical power incident on the chip. Because of the poor optical efficiency of our disks, this corresponds to a 100 W laser source. However, by switching to disks that use different materials (e.g. transmissive disks), we can achieve 10 to 100 times greater efficiency reducing the power requirements of the laser source to the 1 to 10 W range.

We have implemented a two-layer heteroassociative memory using the chip and the optical disk. One of the input vectors, $(++\ ++----+-+-+\ -+)$, is displayed in the bank of LEDs at the top of figure 11(*a*). These LEDs are used only for displaying the state of the neurons on the chip. Valid input vectors consist of 8 +1s (on LEDs) and 7 −1s (off LEDs). Each neuron in the second layer acts as a 'grandmother-cell', recognizing a specific input vector and associating it with a specified output vector. The first light pattern striking the photodiodes is chosen such that it connects each grandmother-cell neuron with only those input neurons that are set to −1s by the specified input vector. With the threshold set to find all seven −1s, the output of each grandmother-cell will be −1 if and only if the input corresponds to the specified vector. The LEDs on the left of figure 11(*a*, *b*) show that the second vector (second LED off) has been recognized. To readout the associated vector, the second light pattern connects each grandmother-cell with output neurons that are supposed to be −1s in the specified output vector. By setting the threshold to find a single −1, a grandmother-cell with a −1 output forces the outputs of the appropriate neurons $(--++-+-++--)$ to −1 as shown at the top of figure 11(*b*). Since there are 11 neurons in each layer, we were able to store 11 different heteroassociations using the chip. During



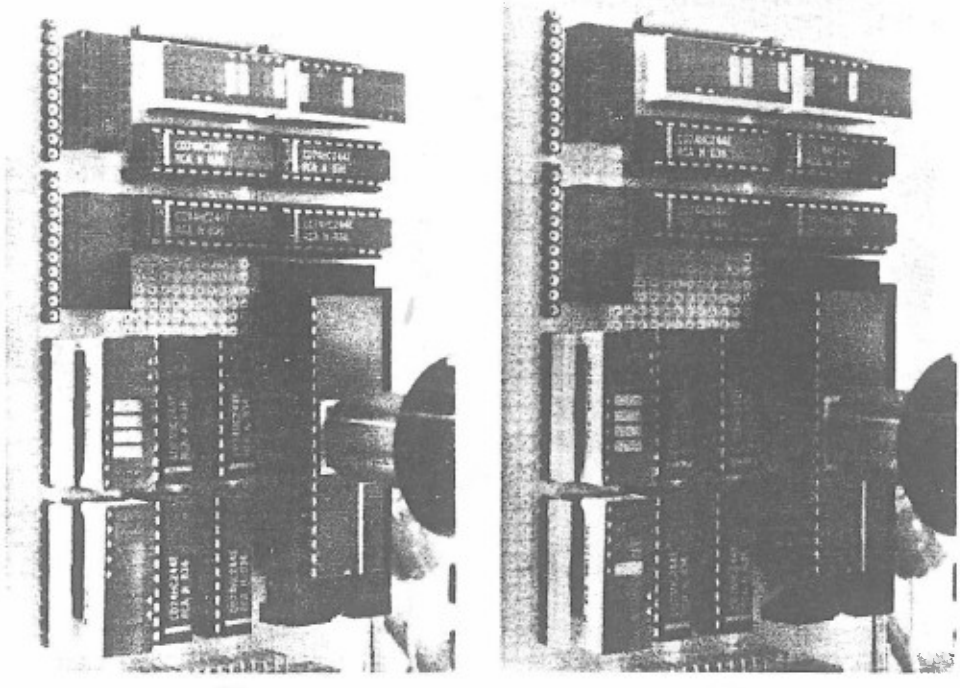
(*a*) (*b*)

Figure 11. Heteroassociative memory: *(a)* first layer, *(b)* second layer.

testing, the network successfully recalled all 11 associations without error.

We have also implemented a two-layer autoassociative memory using the disk and chip with 10 input neurons, 5 neurons in the hidden layer, and 10 output neurons. This time, however, instead of using a grandmother-cell representation, we use a distributed representation in the hidden layer. We train the two-layer network using a combination of two algorithms. The first algorithm [11] uses learning by choice of internal representation and is designed for use with multilayer networks of binary threshold elements; the algorithm applies the perceptron learning rule to find the desired analogue weights in each layer. The second algorithm [12] is a modified version of the perceptron learning rule that finds binary weights. In this hybrid algorithm, we begin with randomly selected binary weights in the first layer. Presenting the training samples at the inputs, the resulting internal representation is tabulated. We then train the second layer with the binary perceptron algorithm by first applying perceptron learning and then thresholding the resulting analogue weights to generate binary weights. If the thresholded binary weights do not produce the correct outputs, we exhaustively try all neighbouring binary weight vectors within a Hamming distance of 1 or 2. If this procedure does not successfully find weights for the second layer, we modify the internal representation by flipping bits and again apply the binary perceptron to the second layer using the new representation. Once we discover an internal representation for which a corresponding binary second-layer weight vector exists, we search for binary first-layer weights that generate the desired internal representation by applying the same binary perceptron algorithm to the first layer. If we fail to find the desired first-layer weights, we record the existing internal representation and again start training the second-layer weights, repeating the above process as required.

Despite the fact that the chip provides unipolar binary weights (0 or 1), we assume the presence of bipolar binary weights (± 1) while training the autoassociative memory since networks with bipolar binary weights have better functionality than those with unipolar binary weights. We can nevertheless ensure that operation of the first layer using unipolar weights is functionally equivalent to that using bipolar weights by requiring that half the 10 input bits be $+1$ and half be -1. (Though the weights are unipolar, the neuron outputs are bipolar.) In this case, we can regard the unipolar weight vector for each neuron to be the sum of a bipolar vector and a bias vector of all $+1$s. The inner product of the bipolar part with the input vector yields the desired result while the inner product of the constant bias with the half-on half-off input goes to zero. We implement bipolar binary weights in the second layer using a dual-rail coding system by taking the five hidden-layer neuron outputs generated by the first



(*a*) (*b*)

Figure 12. Autoassociative memory: (*a*) first layer, (*b*) second layer.

layer and setting five additional hidden-layer neuron outputs to the complements of the first. Each output neuron can then connect to either the output of a hidden-layer neuron or its complement. The difficulty arising from the restriction to unipolar weights leads us to believe that future versions of the optoelectronic neural network chip should implement bipolar connections. Bipolar connections can be provided by the simple addition of a single transistor to each synapse for negative weights connected to a second line providing an inhibitory input to each neuron.

Figures 12(a, b) show the recognition and recall of one of the stored vectors. The input vector ($++--++---+$) is shown at the top of figure 12(a). The left side of figure 12(a) shows the corresponding internal representation ($++++-$) in the hidden layer. The left side of figure 12(b) shows the addition of an externally generated complement to the internal representation ($++++- ----+$). The top of figure 12(b) shows the successful recall of the stored vector. We stored and recalled six vectors in this 10–5–10 network, again without error.

5. Conclusions

Artificial neural systems implement a computational paradigm inspired by biology by use of a large number of simple processing elements with massive interconnectivity between processors. Optics may very likely play a vital role in neural network implementation because of its strengths in communications and memory storage. Optical disks can simultaneously provide a large number of connections and store an entire library of interconnection patterns. We have demonstrated some of these advantages provided by the use of optical disks in two experimental neural network implementations.

Acknowledgements

This research is supported by the Army Research Office and the Defense Advanced Research Projects Agency. Thanks to Robert Snapp for his help in working on the binary multilayer network training algorithm. Alan Yamamura is supported by a fellowship from the Fannie and John Hertz Foundation.

References

[1] Lippmann, R. R., 1987 *IEEE ASSP Magazine*, **4**, 4.
[2] Psaltis, D., Yamamura, A., Neifeld, M. A., and Kobayashi, S., 1989, *Optical Computing 1989 Technical Series Digest*, **9**, 58.
[3] Yamamura, A., Kobayashi, S., Neifeld, M. A., and Psaltis, D., 1990, to be published in *Proc. SPIE, OE-LASE '90.*
[4] Lu, T., Choi, K., Wu, S., Xu, X., and Yu, F. T. S., 1989, *Appl. Optics*, **28**, 4722.
[5] Psaltis, D., Neifeld, M. A., Yamamura, A., and Kobayashi, S., 1990, *Appl. Optics*, **29**, 2038.
[6] Abu-Mostafa, Y. S., and Psaltis, D., 1985, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7, 46.
[7] Le Cun, Y., *et al.*, 1989, *IEEE Communications Magazine*, **27**, 41.
[8] Duda, R., and Hart, P., 1973, *Pattern Classification and Scene Analysis* (New York: John Wiley and Sons).
[9] Poggio, T., and Girosi, F., 1989, *A Theory of Networks for Approximation and Learning*, MIT AI Laboratory and Center for Biological Information Processing, Whitaker College, AI Memo 1140, CBIP paper 31.
[10] Boyd, G. D., 1987 *Appl. Optics*, **26**, 2712.
[11] Grossman, T., Meir, R., and Domany, E., 1989, *Learning by Choice of Internal Representation*, Weizman Institute of Science, Rehovot, Israel
[12] Mok, F. H., 1989, *Binary Correlators for Optical Computing and Pattern Recognition*, PhD thesis, California Institute of Technology.

Accepted for publication in Applied Optics

Optical Implementations of Radial Basis Classifiers

Mark A. Neifeld

Department of Electrical and Computer Engineering

University of Arizona

Tucson, AZ. 85721

Demetri Psaltis

Department of Electrical Engineering

California Institute of Technology

Pasadena, CA. 91125

ABSTRACT

We describe two optical systems based on the radial basis function approach to pattern classification. An optical disk based system for handwritten character recognition is demonstrated. Tbe optical system computes the Euclidean distance between an unknown input and 650 stored patterns at a *demonstrated* rate of 26,000 pattern comparisons per second. The ultimate performance of this system is limited by optical disk resolution to 10^{11} binary operations per second. An adaptive system is also presented which facilitates on-line learning and provides additional robustness.

KEYWORDS: neural networks, radial basis functions, pattern recognition, optical disk.

I. Introduction

We describe two optical architectures for the realization of distance-based classifiers and in particular, radial basis function classifiers. The first is an optical disk based implementation. The 2-D storage format of the optical disk makes parallel access to data an attractive possibility. The optical disk can be thought of as a computer addressed, 2-D binary spatial light modulator or storage medium with a space bandwidth product

of 10^{10} pixels. A number of potential applications which take advantage of these characteristics exist and have been discussed elsewhere in the literature[1,2]. An optical disk based implementation of radial basis classifiers is quite natural owing to the large storage requirements typical of such pattern recognition algorithms. In the system described here, the optical disk based radial basis function classifier is demonstrated as a handwritten character recognition system. The second architecture is a parallel adaptive neural network which facilitates on-line learning and offers added robustness to noise and optical system imperfections.

II. Radial Basis Functions

The RBF approach to pattern recognition differs from neural networks which are based on supervised output error driven learning algorithms like Back Error Propagation (BEP) in a number of respects[3,4]. It is typical for RBF based systems to incur very short learning times while requiring rather large network realizations. This approach therefore, bears some similarity to memory intensive sample based systems such as K-Nearest Neighbor (KNN) classifiers[5]. In such systems, learning time and learning algorithm complexity are traded for classification time and memory requirements. The motivation for using a RBF network to perform pattern recognition tasks comes from the relatively well established mathematical framework associated with regularization theory and hypersurface reconstruction[6]. In hypersurface reconstruction the problem is to construct an approximate function $\hat{f}(\underline{w}, \underline{x})$, which takes a vector $\underline{x}$ into a prescribed output $f(\underline{x})$. The vector $\underline{w}$ is a parameter vector used to tune the estimate $\hat{f}$. For simplicity, we will consider only one dimensional outputs. In order to construct $\hat{f}$ a set of training samples taken from $f(\underline{x})$ (ie. the underlying hypersurface to be approximated) is provided $\{\underline{x}_i \rightarrow f(\underline{x}_i); i = 1, ..., M\}$. The problem then reduces to the choice of the *form* of $\hat{f}$ and the appropriate parameters $\underline{w}$, such that $\hat{f}(\underline{w}, \underline{x}_i) = f(\underline{x}_i)$ for $i = 1, ..., M$. This problem is identical to the pattern recognition problem where one is given a set of training patterns and is asked to find a classifier $\hat{f}$ with the appropriate parameters $\underline{w}$, such that the resulting machine classifies the training set correctly. In both cases we desire that future samples be mapped correctly and that the system behave well in the presence of noise. In order to obtain these desirable characteristics in hypersurface reconstruction, a criterion of smoothness is often placed on the estimator $\hat{f}$. The RBF approach may be

derived as the optimal solution to the regularized problem for a specific smoothing operator.[7] The RBF solution defines an approximating function $\hat{f}(\underline{w}, \underline{x})$ as a weighted sum of radially symmetric basis functions in $\Re^N$. Given a training set $X = \{\underline{x}^i, f(\underline{x}^i); i = 1, ..., M\}$ comprising a set of M points $\{\underline{x}^i \in \Re^N; i = 1, ..., M\}$ and the values of the unknown function $f(\underline{x})$ at those points, the RBF approach specifies an estimator as

$$\hat{f}(\underline{w}, \underline{x}) = \sum_{i=1}^{\bar{M}} a_i exp(-|\underline{x} - \underline{t}^i|^2/\sigma_i^2), \tag{1}$$

where the "centers" or "templates" $\{\underline{t}^i\}$, the "widths" $\{\sigma_i\}$, and the weights $\{a_i\}$ comprise the parameter vector $\underline{w} = \{\underline{t}^i, \sigma_i, a_i : i = 1, ..., \bar{M}\}$, and are determined from the training set.

The RBF classifier seeks to approximate the underlying function as a sum of gaussian "humps". According to the above expression, $\hat{f}$ comprises $\bar{M}$ of these bumps each centered at $\underline{t}^i$ with width σ_i and weighted by a_i to form the final output. We may estimate the parameters $\underline{w}$ from the training set such that $\hat{f}(\underline{x}_i) \approx f(\underline{x}_i)$ using any number of supervised and/or unsupervised algorithms[3,7].

The RBF approach may also he considered as a neural network architecture as shown in figure 1. We define the RBF unit in Fig. 1a as a "neuron" with response given by

$$y_i = exp(-|\underline{x} - \underline{t}^i|^2/\sigma_i^2),$$

where $\underline{t}^i$ is called the "neuron center" and σ_i the "neuron width." These units are depicted in the second layer of figure 1b. The output layer of the RBF network consists of a single linear unit whose output is simply the weighted sum of its inputs. The overall network mapping then is Eq.(1) as desired. In figure 2a we show a RBF network for estimating a function of two input variables and in figure 2h we depict an example of an input space configuration of the mapping induced by such a network. The small disks in fig. 2b represent the training samples and the broken circles represent the e^{-1} contours of the four gaussian hasis functions used to construct the RBF network. As a specific example of training such a network we utilized a k-means algorithm with k=4 to determine the centers of the basis functions[8]. This procedure results in determination of the four centers shown as large asterisks in the figure. In order to determine the widths associated with each center, a KNN algorithm was used. The five nearest neighbors to each center were chosen and the average of these five distances was used as σ_i for the associated bump. Note that these procedures result in the determination of the centers $\underline{t}_i$ and the widths σ_i in a completely unsupervised fashion. In this way

the first layer of a RBF network may be trained without using an error driven procedure thereby reducing training time. Training of the output layer can be accomplished through the use of either a mean squared error minimization procedure (e.g., adaline) or a relatively simple perceptron learning algorithm[9].

III. RBF Based Handwritten Character Recognition

In this section we describe the implementation of the RBF classifier trained to solve a handwritten character recognition problem. We will consider the 10 class problem of identifying handwritten digits 0-9. Using a SUN3/60 workstation, several authors were asked to draw the numerals 0-9 on a 16×16 grid. The resulting database of 950 images (95 per class) was randomly separated into a 300 element testing set and a 650 element training set which will form our reference library. Examples of characters from the training and testing sets are shown in figure 3.

In order to provide shift and scale invariance we first preprocessed both training and testing sets so that each 16×16 image was centered (by repositioning each character within the 16×16 grid such that the number of blank rows/columns of pixels is the same on either side of the character) and scaled to a 10×10 window (by stretching each character such that it's maximum extent is 10 pixels). Following this preprocessing, the 10×10 pixel input field is ***unrastered*** to form a 100 bit binary vector and each such vector $\underline{t}^t$ corresponding to each of the 650 preprocessed training or reference images, is stored on the optical disk as a *radial* line. For each vector $\underline{t}^t$, we also store its complement $\overline{\underline{t}^t}$ in the adjacent position. This method of encoding allows us to simulate bipolar templates on our disks which can store binary, unipolar reflectivity values. The pixel size in this experiment was chosen to be 177 tracks by 116 pixels along track. Track to track spacing is $1.5\mu m$ and pixel separation is approximately $1.0\mu m$. This storage scheme allows us to record 1376 templates per disk.

The architecture we have implemented is shown in Figure 4. The preprocessed 100 bit binary vector $\underline{x}$, is presented to the system shown in Fig. 4 and the first layer of RBF units compute the RBF projections $y_t = exp(-|\underline{x} - \underline{t}^t|^2/\sigma_t^2)$. We have chosen to use as RBF centers $\{\underline{t}^t\}$, all 650 reference images of the training set. This choice of centers also facilitated an earlier KNN based handwritten character recognition system

wbicb has been reported elsewbere[10]. After tbe RBF projections are calculated in the middle layer, this 650 dimensional intermediate representation is then transformed using the interconnection matrix $\underline{\underline{W}}$ to arrive at a 10 dimensional output representation as shown. Eacb output neuron corresponds to one of the ten classes and a winner take all network tben performs the classification. Since we bave cbosen to use the entire 650 template training set as RBF centers, tbe only iterative learning required for tbe first layer of tbis network is for the widtbs $\{\sigma_i\}$. Tbe second layer of course must also be trained to perform the desired classification on the resulting RBF representations.

There are many potential training algorithms for $\{\sigma_i\}$ and $\underline{\underline{W}}$. The most successful algorithm we found for computing the widths $\{\sigma_i\}$, was to make σ_i proportional to the distance between template $\underline{t}^i$ and its nearest neighbor. That is

$$\sigma_i = \tilde{\sigma} \min_{j \neq i} |\underline{t}^j - \underline{t}^i|,$$

where the proportionality constant $\tilde{\sigma}$ is selected *a priori*. Training of the output layer was most successful when $\underline{\underline{W}}$ was initialized with a binary address algorithm and then trained using the perceptron learning algorithm. The binary address algorithm does not require specific knowledge of the intermediate representations generated during training, it only requires knowledge of the class assignment of each of the 650 RBF centers. This reduces second layer computation time and improves network performance. The binary address algorithm defines tbe initial $\underline{\underline{W}}$ as

$$w_{ij} = \begin{cases} 1 & if \quad \underline{t}^j \in \Omega_i \\ -1 & otherwise. \end{cases}$$

Following this initialization, the perceptron algoritbm is used to incorporate detailed knowledge of the training representations into the output layer weights.

Using these procedures for training the RBF network, we have in computer simulation, a best RBF performance of 89% as shown in Table 1. Altbough the trend witb increasing $\tilde{\sigma}$ is an improvement in network performance, in general we found that the broader the basis functions, the longer the perceptron algorithm will take to converge. For this reason, table 1 does not contain any entries for $\tilde{\sigma} > 1.2$. We note here that the best RBF network performance of 89% is substantially better than the best KNN system performance of 83% using the same template library. This performance can also be compared with a single layer of 10

neurons, each trained with the perceptron algorithm using the 650 image reference library. The recognition rate in this case is 75% on the 300 element testing set. In general, we would expect an improvement in RBF network performance with variable centers where $\vec{M}$, $\underline{t}^i$, and $\vec{\sigma}_i$ are all optimized. This case was not studied here as we are primarily interested in the performance of the optical implementation.

In Figure 5 we show the RBF widths computed using the procedure described above. Each row in the figure represents the widths associated with centers in a single class. There are therefore 65 blocks per row and 10 rows in Figure 5. Each block in the figure is a grey scale coding of the width associated with the corresponding template with dark = 0 width. Using this encoding, each row of the figure corresponds to the values of σ_i for templates in a single class. It is interesting to notice that the second row in Figure 5 corresponding to handwritten ones, is particularly dark indicating that these vectors tend to be well clustered or in general located close to other vectors. Also in Figure 5 we can see that the width associated with one particular template representing a handwritten six, is quite broad indicating that this vector is basically isolated in the input space. Using the same display format as in Figure 5, Figure 6 shows the ten weight vectors of the second layer generated for the 'best' RBF network. The single bright row in each weight vector indicates that the weight vector is tuned to intermediate representations from essentially one class.

IV. Optical RBF Classifier

A schematic of the optical system used to compute the distance between an unknown preprocessed input image and each template stored on the disk in the format described above, is shown in figure 7. In this architecture, an Epson LCTV is used as a 1-D SLM to present the unrastered input character to the system. An image of the input vector is formed as a radial line on the disk as shown, and the total diffracted intensity is collected by the output lens and measured using a Photodyne 1500XP detector. The detector output represents the inner product between the input vector and the illuminated template vector. The postprocessing system for this experiment consists of two parts. First, a sample/hold (S/H) circuit is used to detect the peaks of the raw detector output. The amplitudes of these peaks represent the desired inner products. The S/H circuit is clocked by a signal which is phase locked to the sector markers that are recorded

on tbe disk whicb appear as 32 brigbt radial lines and provide a strong diffracted signal. Tbe second stage of postprocessing consists of an A/D converter board in an IBM PC followed by software which implements the nonlinearity of the second layer and computes tbe final output.

Tbe 650 reference images were preprocessed as described above and stored on tbe disk along witb tbeir complements, as 100 bit binary vectors. Using a disk rotation rate of 20Hz, tbese 1300 vectors were processed at a rate of 26,000 inner products per second equivalent to 2,600,000 binary operations per second. It should be pointed out tbat this relatively slow processing speed arises from a severe under-utilization of disk capacity. In this experiment a large pixel size was used (177 tracks by 116 pixels accross track) in order to provide alignment simplicity. A system which utilized the minimum disk resolution of roughly 1μm pixels together with a disk rotation rate of 100Hz, would achieve an inner-product rate of 10^7 per second corresponding to a raw processing rate of 10^{11} binary operations per second. The 300 testing images were preprocessed as described in Section III, and stored in an IBM PC which drove the LCTV and provided input vectors to the system. An example of the raw detector output for the *all ones* input vector is shown in figure 8. The two tallest peaks in this trace correspond to sector markers on the disk and represent the inner product between the *all ones* vector and itself. From this data we can calculate the effective brightness per input pixel as measured at the detector as 0.6nW. This value is in good agreement with the known optical losses in the system. The other peaks in figure 8 provide normalization data which is stored in memory and read out during postprocessing. Tbe PC samples the inner product signal once per peak, averages 4 rotations worth of data (total acquisition time $\approx$0.2s) and computes tbe Euclidean distances from tbe inner products as :

$$|\underline{x} - \underline{t}^i|^2 = |\underline{x}|^2 + |\underline{t}^i|^2 - 2\underline{x} \cdot \underline{t}^i,$$

where $\underline{x}$ is the unknown input image and $\underline{t}^i$ is a stored template. Since our optical system actually measures $\underline{t}^i \cdot \underline{x}$ and $\overline{\underline{t}^i} \cdot \underline{x}$, we may form the distance for binary vectors as :

$$|\underline{x}|^2 = (\underline{x} \cdot \underline{1}) \qquad \underline{1} = (1, 1, \ldots, 1)$$

$$= \underline{x} \cdot (\underline{t}^i + \overline{\underline{t}^i})$$

so that

$$|\underline{x} - \underline{t}^i|^2 = |\underline{t}^i|^2 + \underline{x} \cdot \overline{\underline{t}^i} - \underline{x} \cdot \underline{t}^i.$$

Once again, $|\underline{t}^i|^2$ for $i = 1, ..., 650$ is stored in normalization memory and read out during the postprocessing stage.

The optical disk based inner product calculations are collected by the postprocessing system which computes the required gaussian weighting and simulates the output layer where a classification is made. These postprocessing steps were carried out off line in software for our experiments. The classification rate for the optical RBF system was 83%. This is compared with a recognition rate of 79% using an *optical* KNN network based on the same template data. A comparison between the performances of the optical system and a computer simulation is shown in Table 2. The table entries indicate the number of correct classifications out of 30 for each of the 10 classes 0-9. The various noise sources in the optical system result in a 6% loss of recognition rate. In order to better understand the effect of these imperfections on the RBF network performance, a computer model was constructed which incorporates error sources such as finite contrast, nonuniform illumination profile, detector noise, and quantization noise. Using values for the error variables as measured from the optical apparatus, we found that nonuniformity of the illumination profile was the limiting factor in our experiment. A plot of classification rate vs. log of the $1/e^2$ gaussian profile width is given in figure 9. We can see from Fig. 9 that for the measured profile parameter of 1.8, the expected recognition rate drops to 86%. This rate then is the noise limited optical system performance and is close to the experimentally demonstrated 83%. The cumulative effect of these errors can be measured a second way, directly from the distance calculations. In figure 10 we show the 650 distances computed for a single input image (a handwritten 3) using both the ideal computer simulation and the optical system. From the figure we see that there is a substantial variation between these two plots. This variation can be quantified by computing the RMS distance error over the entire testing set as :

$$\Delta D_{RMS} = \frac{\sqrt{\frac{1}{M}\sum_{i=1}^{M}(d_i^{Opt} - d_i^{Sim})^2}}{\frac{1}{M}\sum_{i=1}^{M} d_i^{Sim}},$$

where d_i^{Sim} and d_i^{Opt} are the Euclidean distances between the 300 input images and the 650 templates calculated from simulation and the optical system respectively. There are M=195,000 such measurements in our case. For the results presented here, the RMS distance error was found to be ΔD_{RMS} =28.5% . Although this error is quite large, the recognition rate obtained using the optical system is in satisfactory agreement with the expected rate attesting to the robustness of the RBF approach.

V. Parallel Optical Distance Computation

In order to provide additional robustness to the optical system as well as increase the computation speed, a parallel non-disk implementation is proposed. We may observe from the previous discussion that the hardware implementation of a RBF network comprises two primary components: the subsystem to compute M parallel Euclidean distances and the basis function evaluation subsystem. Shown in Figure 11 is an optical system which can be used to realize the required parallel distance computation for the case of binary vectors. A similar system can be used to compute the distances for continuous valued vectors; however, we will concentrate on the binary system for now. In Figure 11, an N-dimensional binary vector $\underline{x}$ is represented as a vertical intensity array in the input plane and each center $\underline{t}^i$ is stored in a vertical column of the $\underline{t}$ transparency shown. This system is "dual rail" since it requires $\underline{x}$, $\underline{t}$ and their complements $\overline{\underline{x}}$, $\overline{\underline{t}}$ respectively. We now show that using this representation, the distance computation may be performed entirely optically.

Given an input $\underline{x}$ and a center $\underline{t}^i$, we can write the Euclidean distance between these two vectors as

$$\begin{aligned} d^i &= |\underline{x} - \underline{t}^i|^2, \\ &= \sum_{j=1}^{N} (x_j - t_j^i)^2, \\ &= \sum_{j=1}^{N} d_j^i. \end{aligned}$$

We can further write the component-wise distances d_j^i in the binary case, as the exclusive or (XOR) of the component bits. That is

$$\overline{d_j^i} = x_j t_j^i + \overline{x_j}\,\overline{t_j^i}.$$

Writing d_j^i in this complement form makes the optical realization more clear. Returning to the system of Figure 11, light from the $\underline{x}$ spatial light modulator (SLM) is collimated in the x-direction and imaged in the y-direction so that immediately to the right of the transparency $\underline{t}$, the component-wise product is formed between the input and all of the centers. That is, we generate the array $\{x_j t_j^i; i = 1, ..., M; j = 1, ..., N\}$. Similarly, in the lower arm of the system the complement array $\{\overline{x_j}\,\overline{t_j^i}; i = 1, ..., M; j = 1, ..., N\}$ is formed and these two arrays are simultaneously imaged onto a contrast reversing SLM . This superposition combined

with the contrast reversal yields the desired component distances d_j^i to the right of the contrast reversing SLM. A good candidate for this contrast reversal SLM is the optically addressed FLC SLM.[11]

Returning once again to Figure 11, after the bitwise XORs are computed as described above, the desired array of distances is obtained by summing in the y-direction using the cylindrical lens shown. A simple 1D SLM can be used to represent the desired widths so that immediately to the right of the output plane shown we obtain the desired terms $\{|\underline{x} - \underline{t}^i|^2/\sigma_i^2; i = 1, ..., M\}$.

Although the system described above operates on 1D arrays of data, a 2D version which is better suited to operating on image data is also possible. This 2D extension is straightforward and involves the use of lenslet arrays for accessing the spatially multiplexed template images stored in the $\underline{\underline{t}}$ and $\underline{\underline{\bar{t}}}$ planes. The details of the 2D system as well as those of an extended 1D system capable of operating on continuous valued vectors are the subject of another paper and will not be discussed here; however, in order to indicate the expected performance limitations of these systems we consider the 2D binary version. If we assume that the inputs to our system are 100X100 pixel images then the number of centers in the RBF network will be limited by the space bandwidth product (SBP) of the optical system. Realization of 900 centers will require a template mask with 3000X3000 pixels which in turn demands an optical system with SBP=$9X10^6$. This would be the largest feasible implementation. Notice that although the contrast reversal plane must have large SBP, the output plane requires only SBP equal to the number of centers. This is an attractive characteristic of the present system since on-line learning will require a programmable SLM in this plane for σ_i adaptation. In the present example, this SLM would be required to have only 30X30 pixels.

VI. Parallel Basis Function Evaluation

Having defined the optical distance computer in the previous section we turn our attention to the second primary component of the optical RBF implementation. This component performs the basis function evaluation. Notice that the basis function evaluation requires only *point* operations in the plane of distances. Further, if we consider the case of a single output neuron (i.e., $f(\underline{x}) : \Re^N \rightarrow \Re^1$) then to complete the RBF computation after the distance computer requires *only* point operations followed by a global sum. With this

observation in mind we propose the optoelectronic postprocessing chip shown in Figure 12.

The chip shown consists of an array of modules each module comprising a photodetector to detect the output of the optical distance computer, analog multipliers to realize the required width and output weighting and an exponentiation unit to realize the hasis function evaluation. The output of each such module is summed on a common line to generate the network response. We should note here that all the required functions in a module are compactly achievable using analog VLSI or if more precision is required, the photodetector may be followed by an A/D converter and each module could then be implemented in digital electronics. The 2D extension of this postprocessing chip is once again straightforward and since there are no intermodule communication requirements, connectivity issues in the 2D arrangement do not arise. All computation in this postprocessing chip is local excepting the final sum. Also note that this implementation has the flexibility to allow for the realization of a variety of different basis functions as well as supporting a useful on-line learning algorithm which will be discussed further in the next section.

The above "all electronic" postprocessing chip is particularly well suited to the case of a network with only one output. For the case of multiple outputs we have two alternative systems. If the number of outputs is relatively small (≈ 10) then the most attractive alternative is simply the use of multiple postprocessing chips. This approach retains the simplicity and flexibility of the VLSI implementation. Alternatively, if the number of outputs is large, we may consider a hybrid approach wherein each $[e^{-x}]$ box in Figure 12 is followed by a light modulating element allowing the exponentially weighted distances to be read out optically using liquid crystal modulators for example.[12] In this way an efficient, optical implementation of the output layer is facilitated. This approach has the advantage of providing scalability in terms of output units while retaining much of the convenience of the VLSI implementation. In this system, update of the output weights during a learning cycle is done optically, through the use of photorefractive holograms in the output layer.[13]

VII. Learning

The effective implementation of iterative learning algorithms is a common stumbling block in both electronic and optical neural network architectures. In this section we suggest a learning algorithm for

RBF networks which is suitable for implementation using the optoelectronic hardware we have described. Associated with each of the postprocessing stages (i.e., all electronic and hybrid) is an implementation of the on-line learning algorithm. We will describe the all electronic single output implementation here.

Referring to Equation 1 for the RBF network response function, we can define a criterion function for the "goodness" of an RBF network as

$$E = \sum_{i=1}^{M} (\hat{f}(\underline{w}, \underline{x}^i) - f(\underline{x}^i))^2,$$
$$= \sum_{i=1}^{M} E_i^2,$$

where E_i is just the error between the actual and desired network responses in the presence of training vector $\underline{x}^i$. E is just the conventional "sum of squared error" function evaluated over the entire training set. If we assume that the training set T is fixed and that each training vector will be used as a single RBF center as before. then the learning procedure reduces to finding $\{\sigma_i\}$ and $\{a_i\}$ to minimize the error E. A simple gradient decent procedure is a candidate algorithm for the minimization of E. Using this procedure we can write expressions for the update of the network parameters σ_p and a_p in response to the error measured for a single input training vector $\underline{x}^l$. These are

$$\Delta(1/\sigma_p)^2 = -\alpha_\sigma E_l |\underline{x}^l - \underline{t}^p|^2 a_p exp(-|\underline{x}^l - \underline{t}^p|^2/\sigma_p^2), \quad (2)$$

$$\Delta(a_p) = \alpha_a E_l exp(-|\underline{x}^l - \underline{t}^p|^2/\sigma_p^2), \quad (3)$$

where α_σ and α_a are acceleration constants for the width and output weight updates respectively[15]. Notice that these expressions define a "backward error propagation" type of rule for RBF networks. In this learning algorithm however, no special hackward response function is required for the RBF units owing to the fact that the $exp(x)$ function is its own derivative. All signals required to compute the updates defined in Equations 2 and 3 above are present in the *forward* path of the network. Furthermore we observe that all required learning signals are present in the electronic portion of the proposed implementation and that no intermodule communication is necessary. The implication of these observations is that rapid, parallel update of all network parameters can be realized with a simple modification of the postprocessing module presented earlier. In Figure 13 we show a block diagram of the modified postprocessing module. By incorporating the

additional "local" connections shown and adding accumulation registers to the $[(1/\sigma)^2]$ and $[Xa]$ blocks, the on-line parallel RBF learning algorithm can be realized with little increase in overall circuit complexity over the non-adaptive system. In this way a 30X30 element adaptive postprocessing array, capable of facilitating on-line learning in a 900 center RBF network should be possible.

VIII. Conclusions

We have demonstrated an optical system which can implement a RBF pattern classifier. The experimental system achieved a processing rate of 2,600,000 binary operations per second corresponding to the computation of 13,000 Euclidean distances per second. The capability of the optical disk based system is limited by the maximum length of template vectors ($\approx 10^4$ bits), the maximum number of template vectors ($\approx 10^5$) and the maximum disk rotation rate ($\approx$ 100Hz). These upper bounds correspond to a processing rate of $\approx 10^{11}$ binary operations per second.

This system was trained off-line, with the handwritten numerals 0-9 and achieved a recognition rate in computer simulation of 89% on a 300 element testing set. Similar performance (91% recognition rate) was achieved using the same off-line training procedure in an RBF network with 2000 centers, trained on segmented zip code data obtained from the U.S. post office database. The optical disk based 650 center system achieved a recognition rate of 83%. In this work it was found that factors such as nonuniform disk reflectivity, nonuniform illumination and finite contrast were all significant contributors to a 28% RMS error in the optical distance computation. Furthermore, since this large distance error resulted in only a 6% degradation in recognition performance, the RBF approach was seen to be robust in the presence of such errors.

We might expect an on-line learning scheme in which optical system imperfections are present during the learning phase, to provide compensation for those imperfections and result in a recognition rate closer to the simulation value. We have described such an adaptive optical RBF hardware implementation. Combining on-line learning, more careful system design, more powerful learning algorithms to learn the optimal center and width values and a larger hidden layer ($\approx$ 2000 units), the optical system should be able to approach the 91% recognition rate obtained in simulation for the zip-code data.

IX. Acknowledgements

This work is supported by the U.S. Army Research Office and the Defense Advanced Research Projects Agency. The authors would like to thank Seiji Kobayashi, Sony Corporation Subrata Rakshit and Alan Yamamura for assistance with the optical disk recording system.

X. References

[1] D. Psaltis, M. A. Neifeld, A. A. Yamamura, "Image Correlators Using Optical Memory Disks," *Optics Letters*, Vol. 14, 429–431, (1989).

[2] D. Psaltis, M. A. Neifeld, A. A. Yamamura, and S. Kobayashi, "Optical memory disks in optical information processing," *Applied Optics*, Vol. 29, 2038–2057, (1990).

[3] John Moody and Christian Darken, "Fast Learning in Networks of Locally Tuned Processing Units," *Neural Computation*, Vol. 1, 281–294, 1989.

[4] D. E. Rumelhart G. E. Hinton and R. J. Williams, "Learning Internal Representations by Error Propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1.* D. E. Rumelhart and J. L. McClelland, eds., (MIT Press, Cambridge, Mass., 1986), pp. 318-362.

[5] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, Vol. IT-13, 21-27, (1967).

[6] T. Poggio, F. Girosi, "A Theory of Networks for Approximation and Learning," *MIT AI Laboratory and Center for Biological Information Processing Whitaker College*, AI Memo No. 1140, CPIB Paper No. 31, July 1989.

[7] T. Poggio and F. Girosi. "Networks for Approximation and Learning," *Proceedings of IEEE*, Vol. 78, 1481-1495, (1990).

[8] J. MacQueen. "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Math. Stat. and Prob. I.*, L. M. LeCam and J. Neyman, eds. (University of California Press, Berkeley and Los Angeles California, 1967), pp. 281–297.

[9] R. Duda and P. Hart, "Pattern Classification and Scene Analysis," (John Wiley and Sons, Ney York, 1973), Chapter 5, pp. 141–147.

[10] M. A. Neifeld, S. Rakshit, A. A. Yamamura and D. Psaltis, "Optical Disk Implementation of Radial Basis Classifiers." *Proceedings of SPIE 1990 International Symposium on Optical and Optoelectronic Applied Science and Engineering*, Vol. SPIE 1347, No.2, San Diego, Ca. 1990.

[11] D. Jared, K. M. Johnson, and G. Moddel, "Joint Transform Correlator using an Amorphous-Silicon

Ferroelectric Liquid-Crystal Spatial Light Modulator," *Optics Communications*, Vol. 76, 97–102, (1990).

[12] T. J. Drabik and M. A. Handschy, "Silicon VLSI Ferroelectric Liquid-Cystal Technology for Micropower Optoelectronic Computing Devices," *Applied Optics*, Vol. 29, 5220–5223, (1990).

[13] D. Psaltis, D. J. Brady and K. Wagner, "Adaptive Optical Networks using Photorefractive Crystals," *Applied Optics*, Vol. 27, 1752–1759, (1988).

[14] M. A. Neifeld, S. Rakshit and D. Psaltis, "Handwritten Zip Code Recognition using an Optical Radial Basis Function Classifier," *Proceedings of SPIE 1991 Symposium on Applications of Artificial Intelligence and Neural Networks*, Vol. SPIE 1469, No.33, Orlando, Fla. 1991.

XI. Figure Captions

Figure 1 : (a) Definition of RBF units and linear units. (b) General RBF network.

Figure 2 : (a) RBF network for estimating a scalar function of two variables. (b) Example input space configuration resulting from the network of (a).

Figure 3 : Example of handwritten numerals from (top) the training set and (bottom) the testing set used in the optical RBF experiment.

Figure 4 : RBF network for handwritten digit recognition.

Table 1 : Classification results obtained using a 1-nearest neighbor rule for training the RBF widths.

Figure 5 : RBF widths computed using the 1-nearest neighbor rule with $\tilde{\sigma} = 1.2$.

Figure 6 : Second layer weights computed using tbe perceptron algorithm after initialization with the binary address algorithm. The weights for neurons 1-10 appear consecutively from left to right and top to bottom.

Figure 7 : Optical system used to compute the distance between an input and an array of stored templates.

Figure 8 : Example of raw detector output indicating the optically computed inner products.

Table 2 : Performance comparison between optical RBF classifier and simulation. Table entries indicate the number of correct classifications out of 30 for each class.

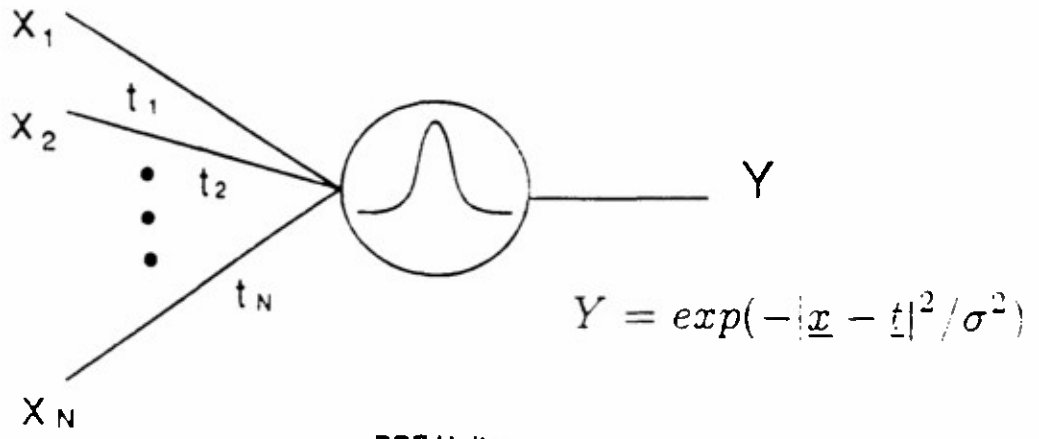
Figure 9 : Predicted recognition rate vs. illumination profile width.

Figure 10 : (a) Experimental and (b) actual distance vs. template number for a single input image (handwritten three number three). Template numbers 195-260 represent the class of handwritten threes.

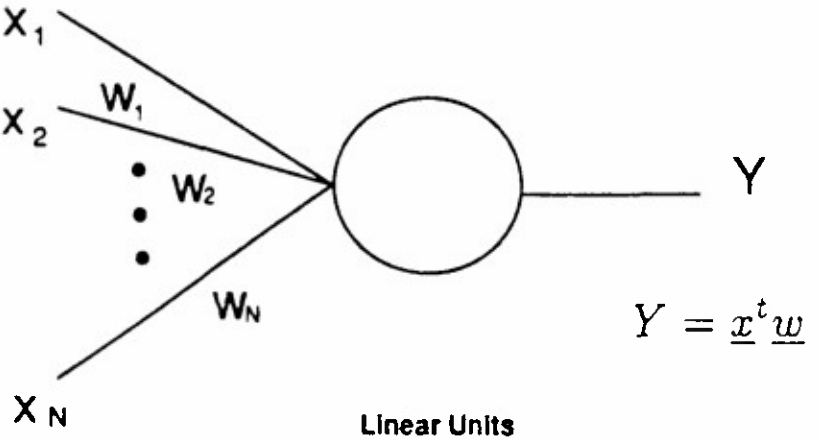
Figure 11 : Parallel optical distance computer.

Figure 12 : Optoelectronic postprocessing chip.

Figure 13 : On-line learning postprocessing module.

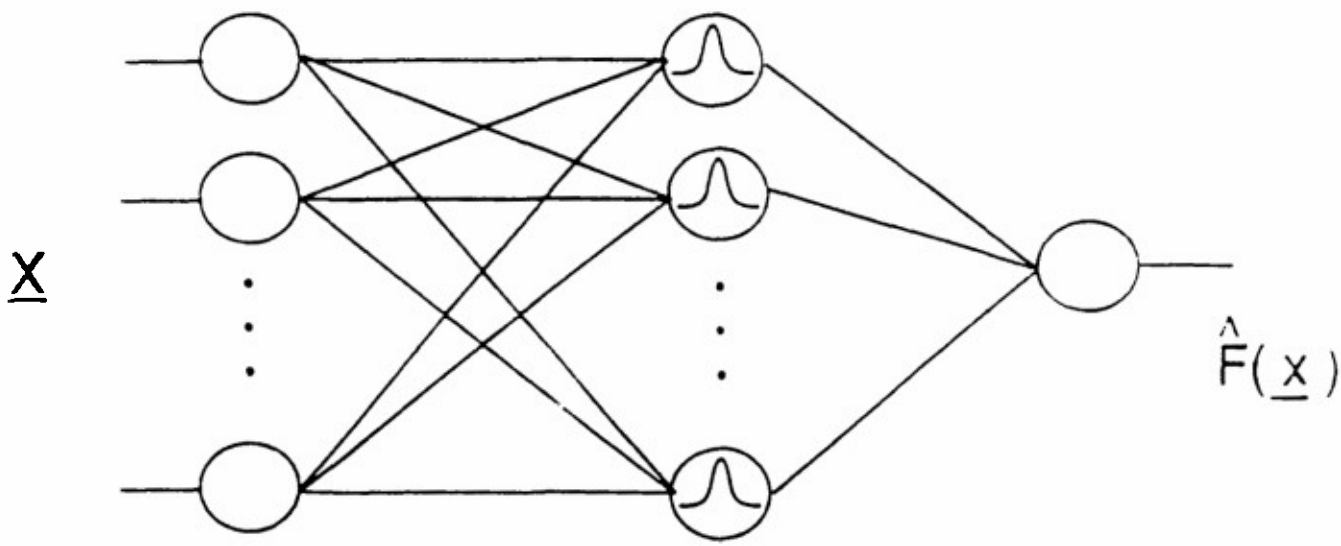


RBF Units



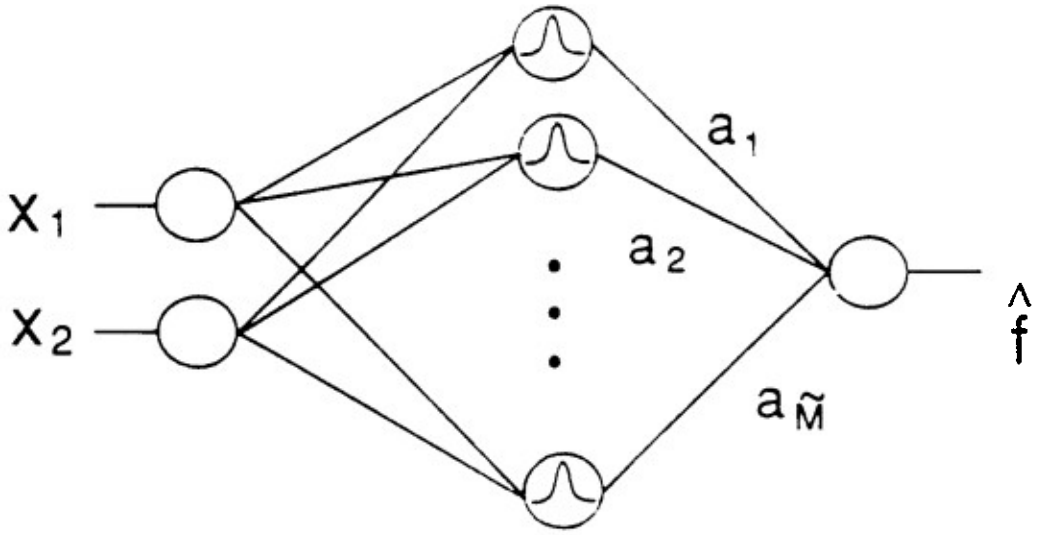
Linear Units

Figure # 12

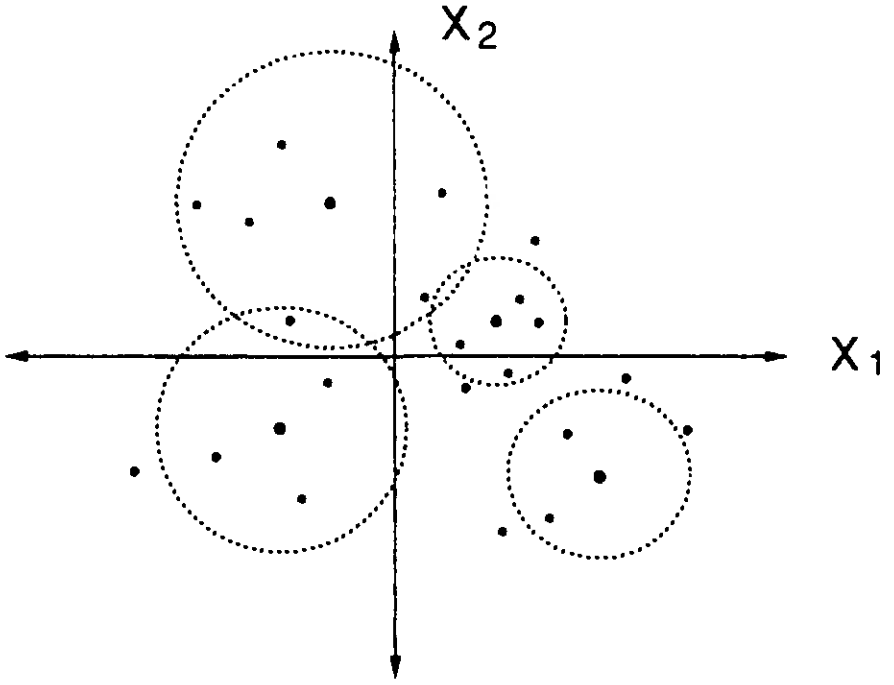


RBF Network

Figure # 16



(a)



(b)

Figure # 2

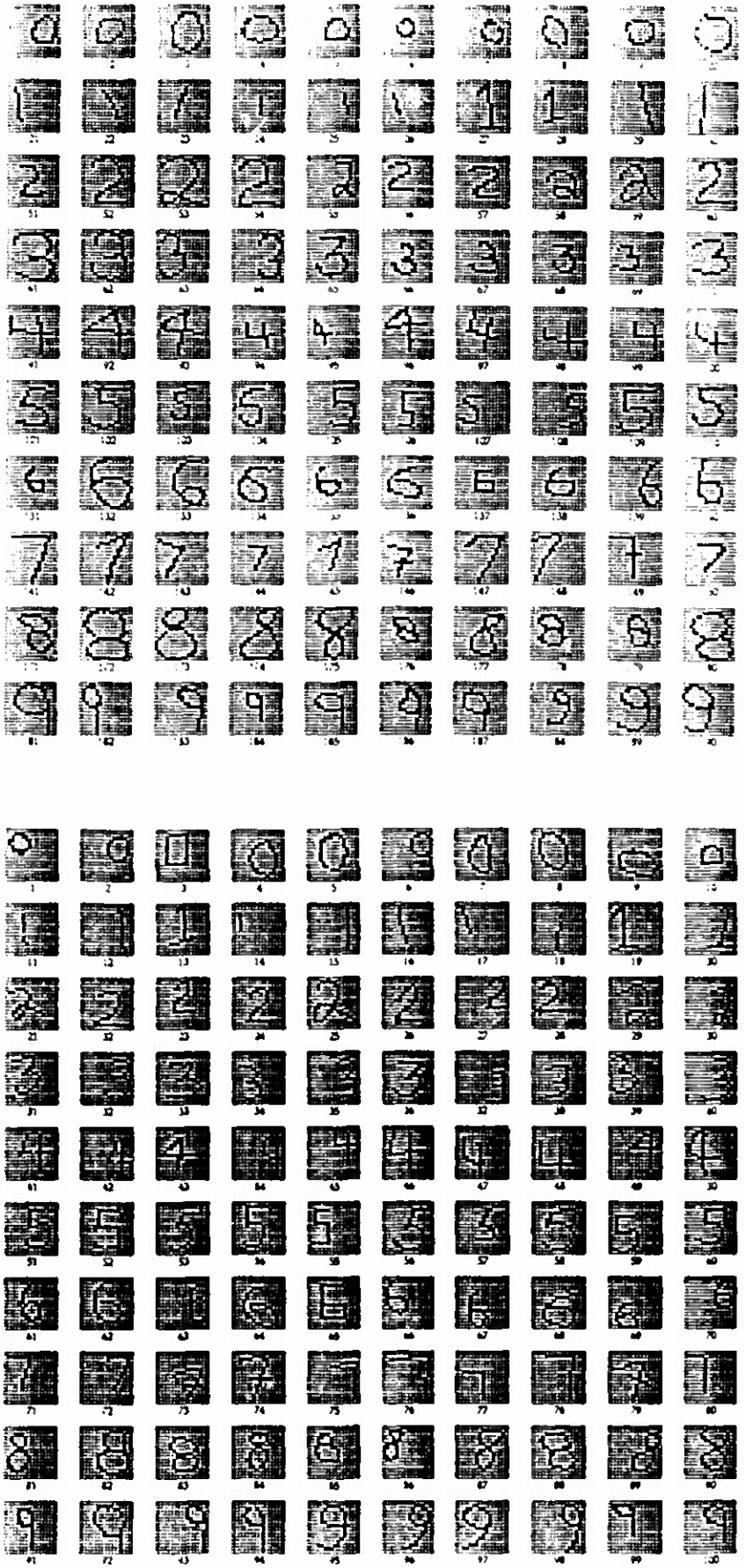


Figure # 3

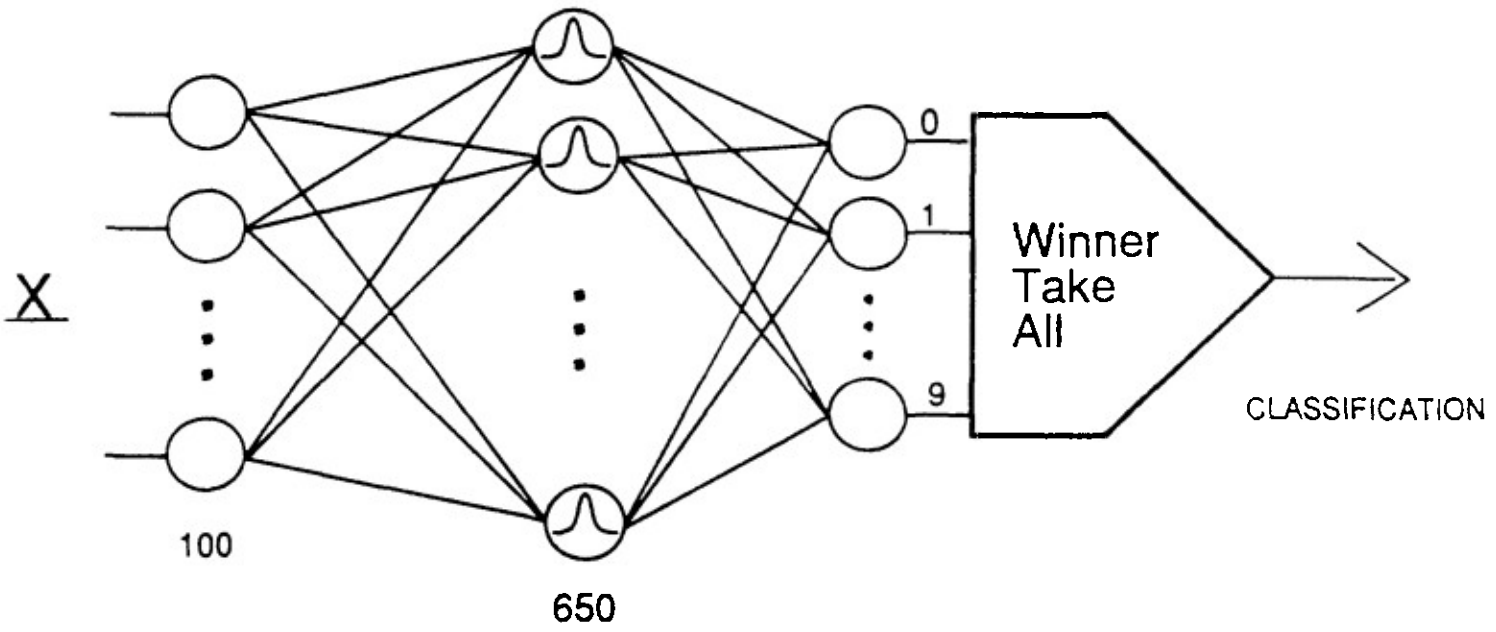


Figure # 4

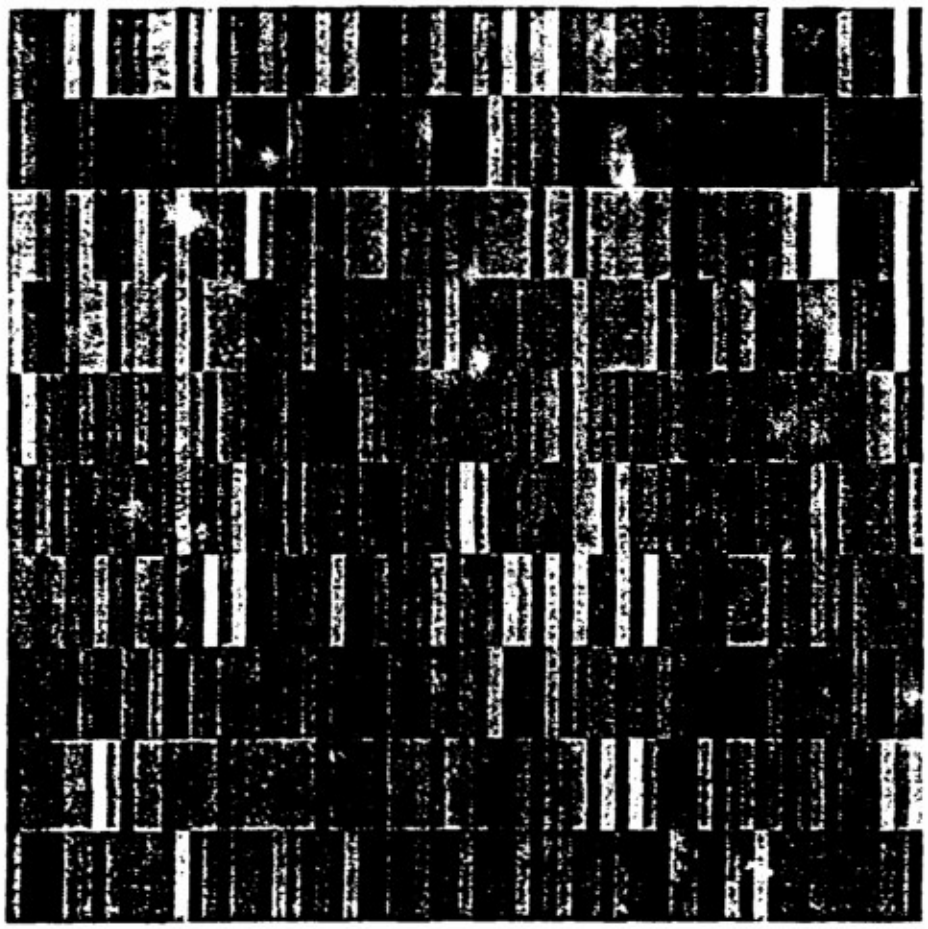


Figure # 5

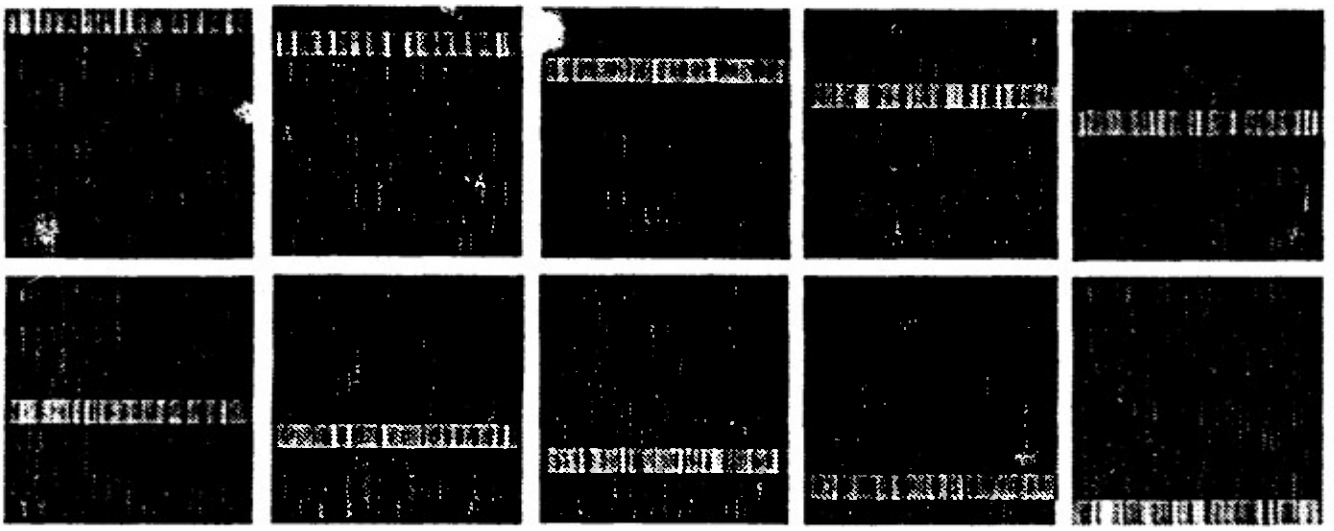


Figure # 6

$\tilde{\sigma}$	TRAINING SET (CORRECT OUT OF 650)	TESTING SET (CORRECT OUT OF 300)
0.5	650	226
0.7	650	257
1.0	650	266
1.2	650	267

TABLE # 1

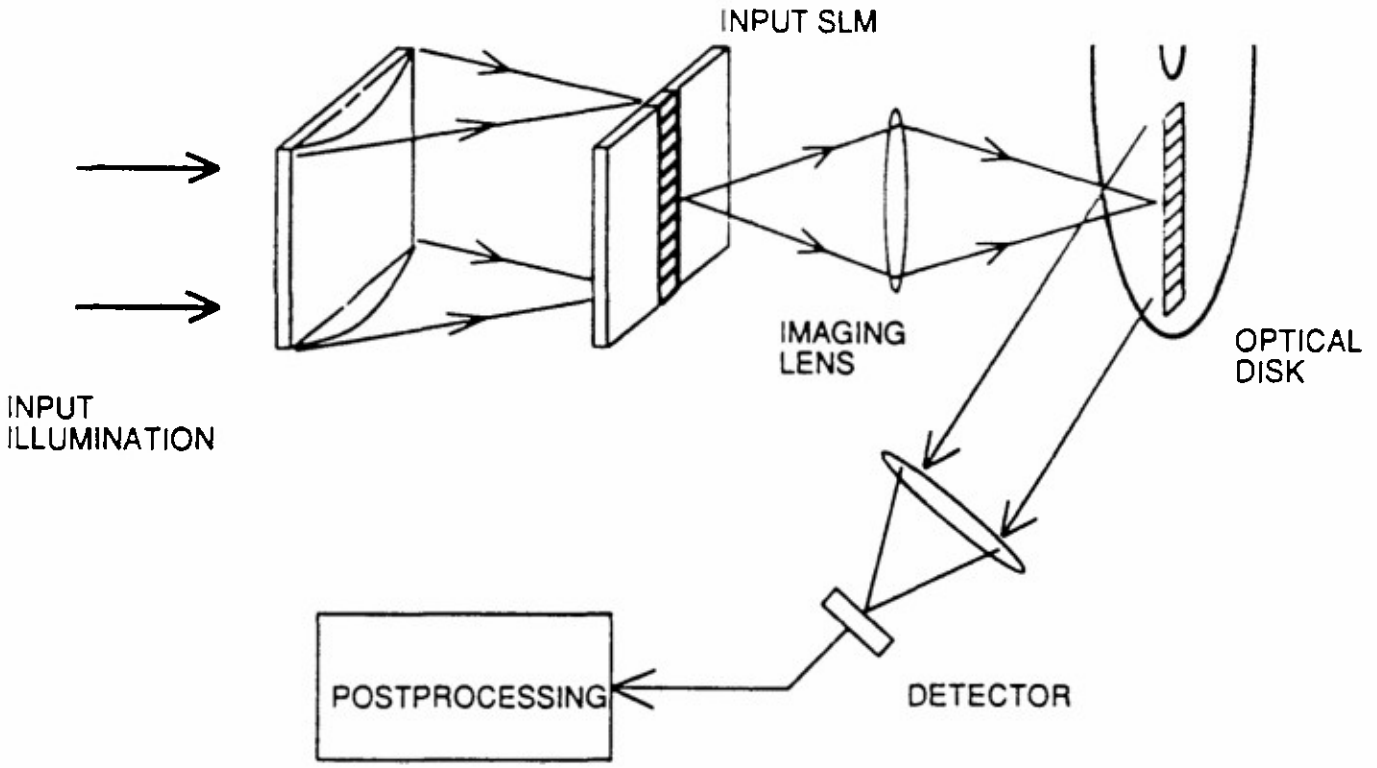

INPUT SLM
IMAGING
LENS
OPTICAL
DISK
INPUT
ILLUMINATION
POSTPROCESSING
DETECTOR

Figure # 7

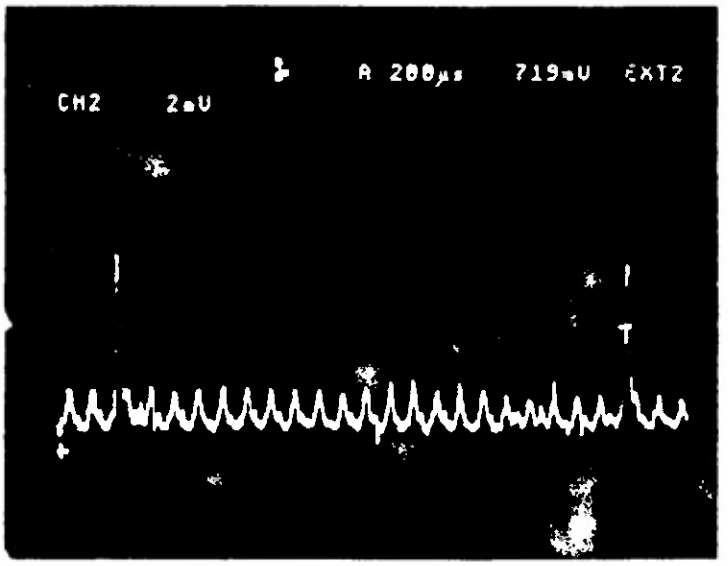


Figure # 8

CLASS	EXPERIMENT	SIMULATION
0	25	29
1	28	29
2	28	28
3	25	27
4	28	23
5	20	25
6	25	24
7	23	28
8	24	25
9	22	29
Total	248	267
Overall Recognition Rate	83%	89%

TABLE # 2

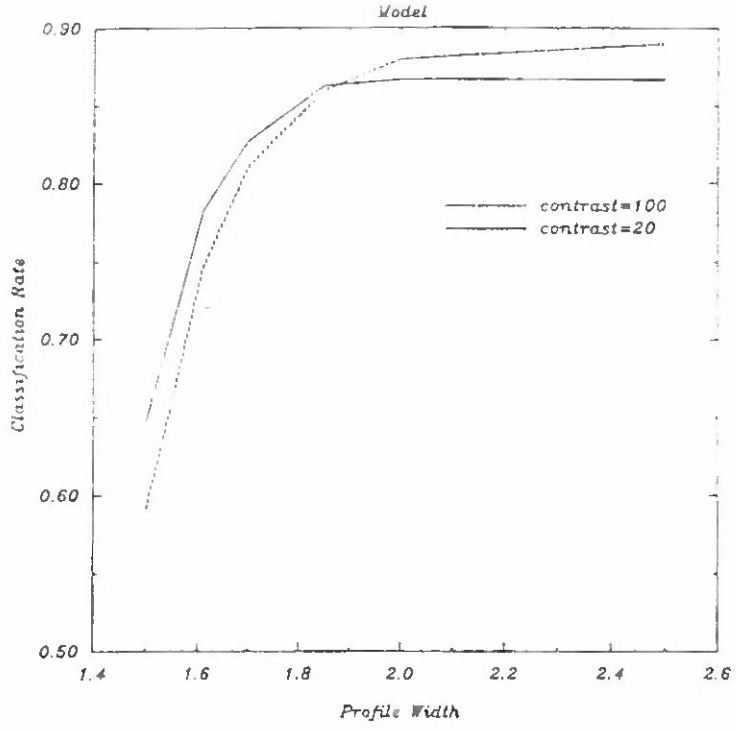
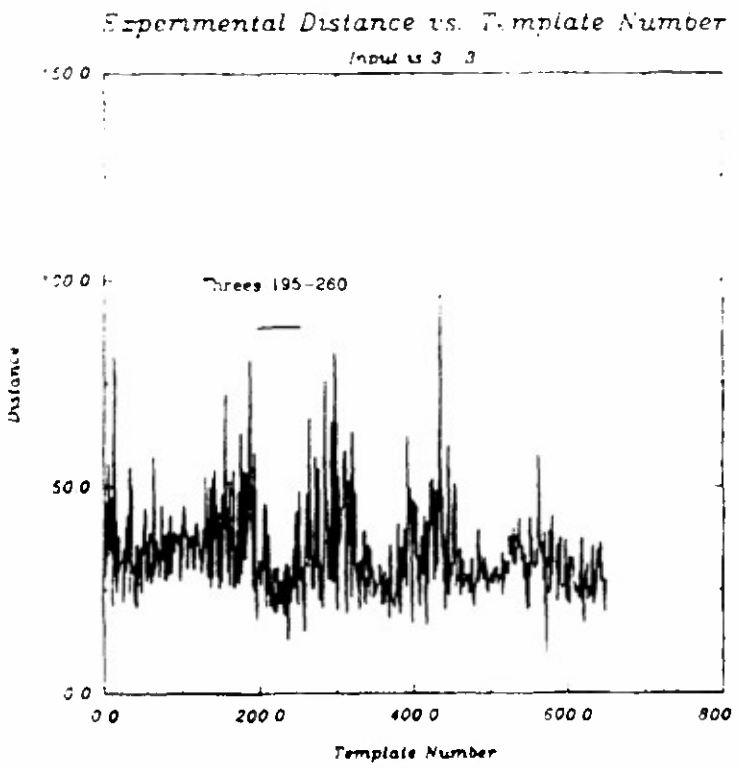
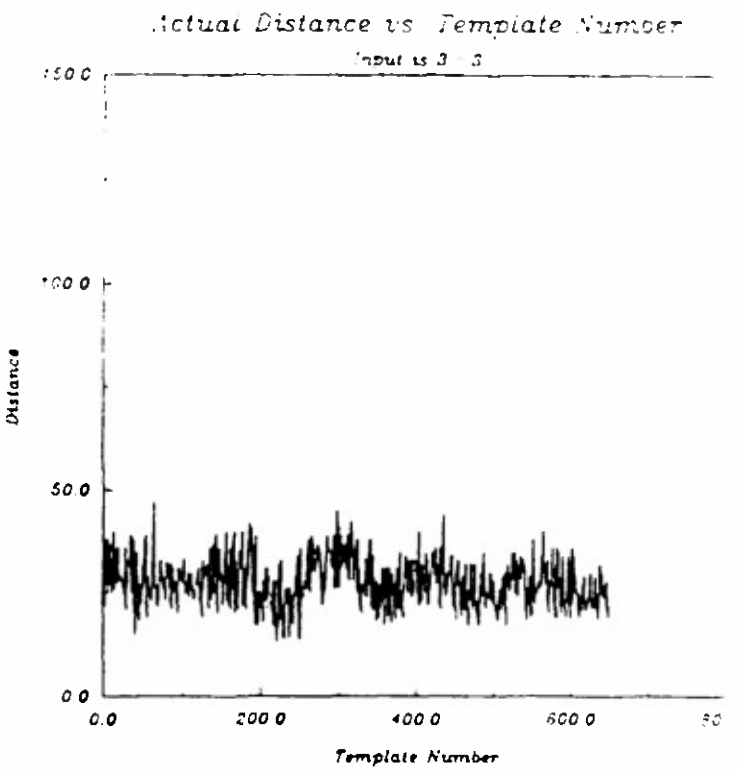

Model
0.90
0.80
0.70
0.60
0.50
Classification Rate
contrast=100
contrast=20
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Profile Width

Figure #9


Experimental Distance vs. Template Number
Input is 3 3
Threes 195-260
Distance
150.0
100.0
50.0
0.0
0.0
200.0
400.0
600.0
800
Template Number

(a)


Actual Distance vs Template Number
Input is 3 3
Distance
150.0
100.0
50.0
0.0
0.0
200.0
400.0
600.0
Template Number

(b)

Figure # 10

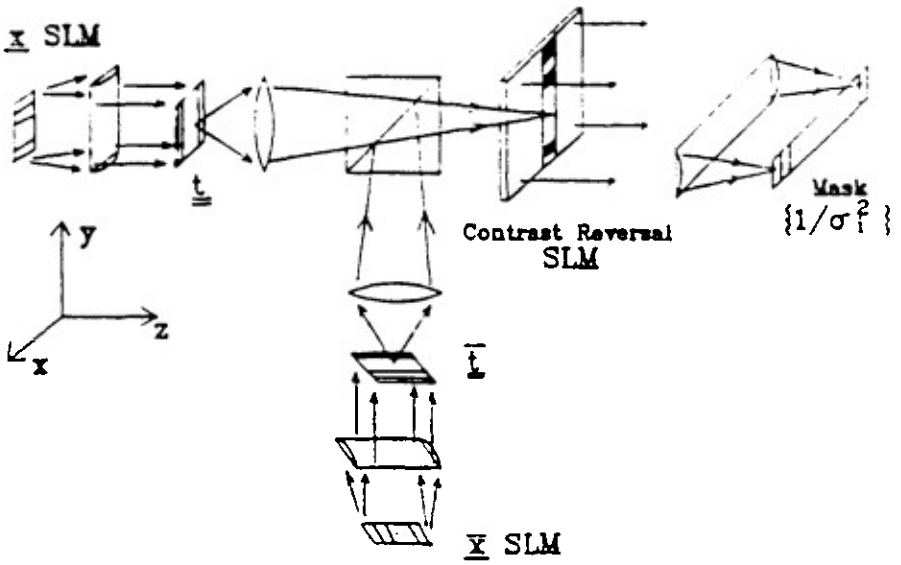


Figure # 11

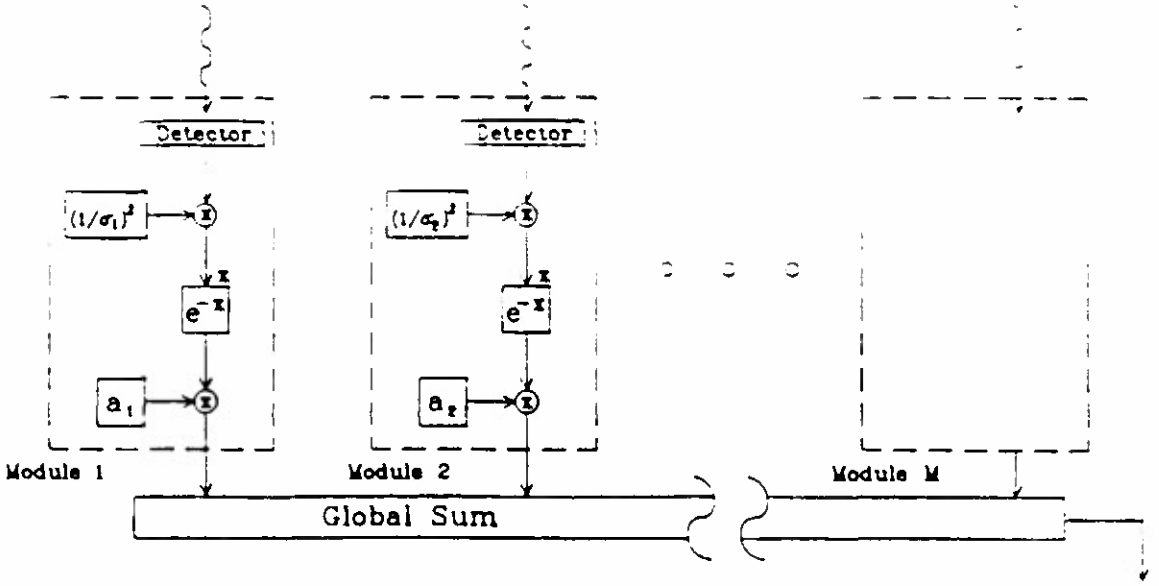

Detector
Detector
$(1/\sigma_1)^2$
$(1/\sigma_2)^2$
x
e^{-x}
a_1
a_2
Module 1
Module 2
Module M
Global Sum

Figure # 12

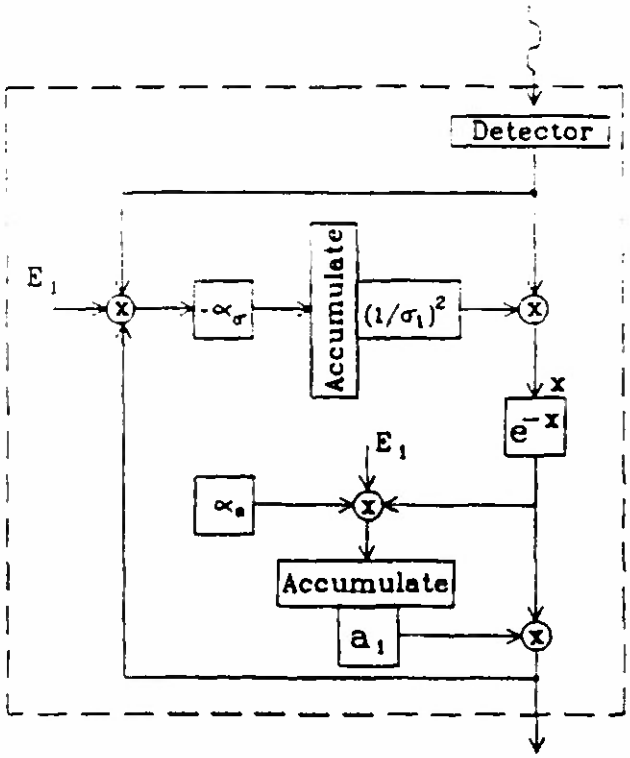


Figure # 13

Image correlators using optical memory disks

Demetri Psaltis, Mark A. Neifeld, and Alan Yamamura

Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125

Received November 14, 1988; accepted February 17, 1989

Image correlators are described and experimentally demonstrated that are implemented using optical memory disks to store a large library of reference images.

Optical correlation for pattern recognition[1] has long been considered a promising application for optical processing. One of the reasons such correlators have not been used in practice is that in many practical applications a single filter is not sufficient to produce reliable recognition. A straightforward solution to this problem is the use of spatial[2] and temporal[3] multiplexing to search through a library of filters. The optical-disk correlator architectures that we describe in this Letter provide an extremely efficient method for performing such a search since they combine in a single device the large memory required for storage of the library of reference images, the spatial light modulator needed to represent the reference in the optical correlator, and the scanning mechanism to search temporally through the library.

The first architecture that we describe is shown in Fig. 1. Each reference image is recorded as a two-dimensional (2-D) computer-generated Fourier-transform hologram (CGH) on the disk. The input image enters the system through the beam splitter, is Fourier transformed by the lens, and illuminates the hologram on the disk. The reflected light contains a term proportional to the product of the transforms of the input and reference images. The same lens retransforms the reflected light, and the correlation is produced at the output plane. An important question in this architecture is whether optical disks are suitable as holographic recording media. Figure 2 shows the diffraction pattern obtained with He–Ne laser light from a write-once disk manufactured by Sony on which we have recorded a 2-D grating. The sharpness of the characteristic diffraction pattern indicates that the glass cover of the disk has sufficient optical quality to allow coherent reconstruction. The rotation of the disk is used to perform a search through images centered at the same radial position on the disk. An auxiliary scanning mechanism is needed in order to position the correlator head in the correct radial position. As the disk rotates it produces a correlation pattern at the output when the transform of the input and the reference hologram on the disk are in alignment.

The above architecture requires storage of the reference images in the form of computer-generated Fourier-transform holograms. A disadvantage of this approach is that it increases the computational overhead for recording the disk. Also, for a Lohmann-type computer-generated hologram the space–bandwidth product required to record the hologram is one hundred times greater than that of the image itself, and the resultant increase in area needed to record each image increases the optical power and phase uniformity requirements. However, in many cases it is necessary only to record reference holograms as binary patterns,[4] in which case each pixel of the image can be directly recorded as a separate spot on the disk. Gray-scale images can be recorded if necessary by the use of some form of area modulation, as is done with video disks, for example.

We discuss two types of architecture that allow the reference images themselves, rather than their Fourier transforms, to be stored on the disk. The first is shown in Fig. 3. The input image goes through the beam splitter and is Fourier transformed by lens L1. A Fourier-transform hologram of the input is recorded in a photorefractive crystal, using a reference beam that is incident from the right. Once the hologram is recorded the input is blocked and the disk is illuminated. Lens L1 takes the Fourier transform of the reference image that is in the field of view of the illuminating beam, and lens L2 transforms the light diffracted by the hologram to produce the correlation at the output plane. If a thick hologram is used, the shift

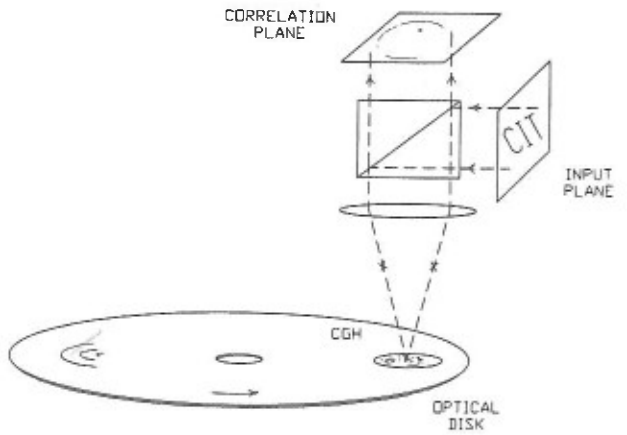


Fig. 1. Optical disk-based Vander Lugt correlator. The reference images are stored on the disk as computer-generated Fourier-transform holograms.

0146-9592/89/090429-03$2.00/0

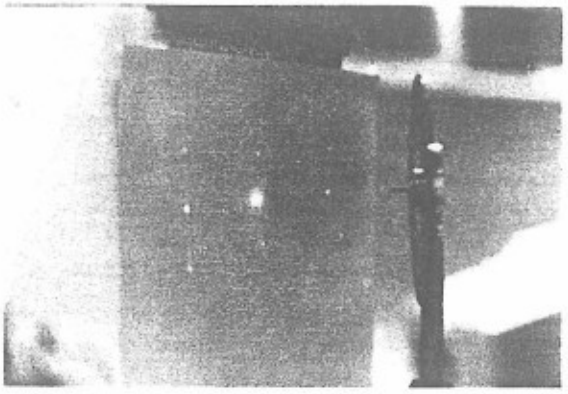


Fig. 2. Diffraction pattern from a 2-D grating recorded on a write-once optical disk.

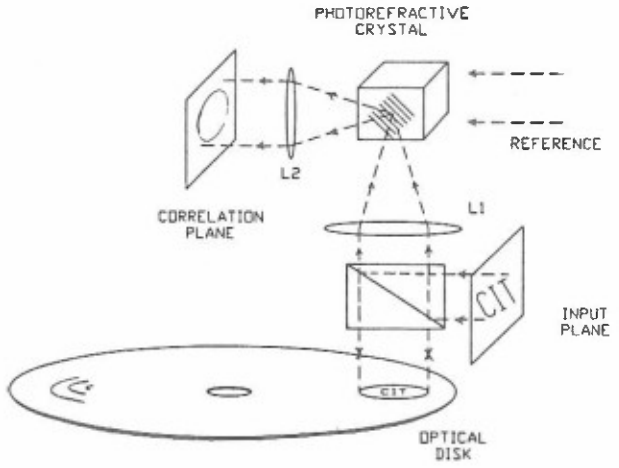


Fig. 3. Optical disk-based correlator implemented by using a photorefractive crystal in the Fourier plane.

invariance is lost in one direction only.[5] Selecting this to be the along-track direction allows the disk rotation to restore 2-D shift invariance. The rotation of the disk is used to search through a library of images in the along-track direction. An advantage of this architecture is that it is invariant to a shift in the position of the recording on the disk. This eliminates the requirement for precise alignment of the correlator head, and in addition a time-delay-and-integrate detector can be used to integrate the traveling correlation pattern at the output, thereby increasing sensitivity. Multiple holograms could be recorded in the crystal to address different radial positions on the disk, or the entire head can be mechanically scanned.

The third architecture that we discuss is shown schematically in Fig. 4. The advantage of this architecture is that it operates on the light intensity, and consequently the requirement for phase uniformity is greatly relaxed.[6] As a result it is possible to implement this architecture with most existing disk systems. This correlator works as follows. The reference images are recorded on the disk, and the input is imaged through a one-dimensional scanning device onto the disk. The scanner can be either an acousto-optic device (AOD) (as shown in Fig. 4) or a rotating mirror. It provides the relative displacement in the radial direction between the input and reference images that is necessary to calculate the correlation function. The disk rotation provides the displacement in the orthogonal direction. The scanner translates the input image completely across the stored reference image each time the disk rotates by a distance equal to a pixel of the reference. The intensity of the light reflected from the disk at any one time is proportional to the product of the input and a shifted version of the reference. The reflected light is collected (integrated) on a single detector that produces as its output a temporal video signal of the 2-D correlation. This system was experimentally demonstrated with a flying-spot acousto-optic scanner in which a chirp signal propagates in the acousto-optic device acting as a traveling lens that scans the diffracted image at a rate equal to the acoustic velocity. This system completes a scan in 30 μsec (the acoustic transit time across the acousto-optic cell), therefore a complete 2-D correlation of an image consisting of 10^2 lines takes approximately 3 msec. A sample of experimental results obtained with a system like that of Fig. 4 is shown in Fig. 5. Figure 5(a) is a photograph of the pattern recorded on a write-once disk (the acronym CIT), and Fig. 5(b) is an oscilloscope trace of the detector signal produced by the optical system of Fig. 4. Figure 5(c) shows the same trace magnified to reveal the individual correlation lines produced by the acousto-optic scanner. The format of the detector signal is similar to a video signal of the 2-D correlation, and it can be displayed in two dimensions by raster scanning the detector output on a 2-D monitor [Fig. 5(d)]. Correlations can be produced with our experimental apparatus at rates up to 1000, 100 × 100 pixel reference images/sec. The optically calculated correlation is in good agreement with the expected autocorrelation function of the CIT pattern. It should be pointed out that since this system operates on intensity, only positive quantities can be represented. Bipolar input and/or reference images can also be represented by adding a bias at the input stage and subtracting it from the output. This tech-

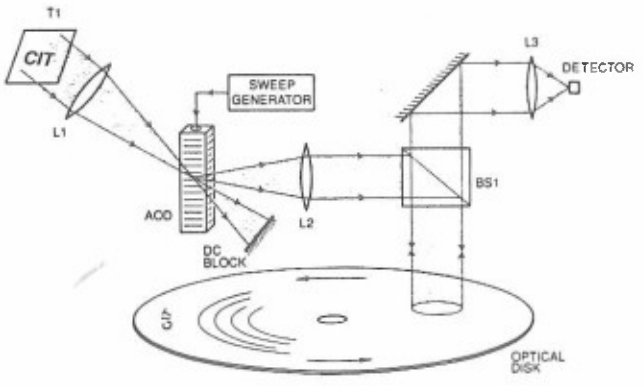


Fig. 4. Incoherent disk-based optical correlator implemented with an acousto-optic scanner. T1, input transparency; L's, lenses; BS1, beam splitter.

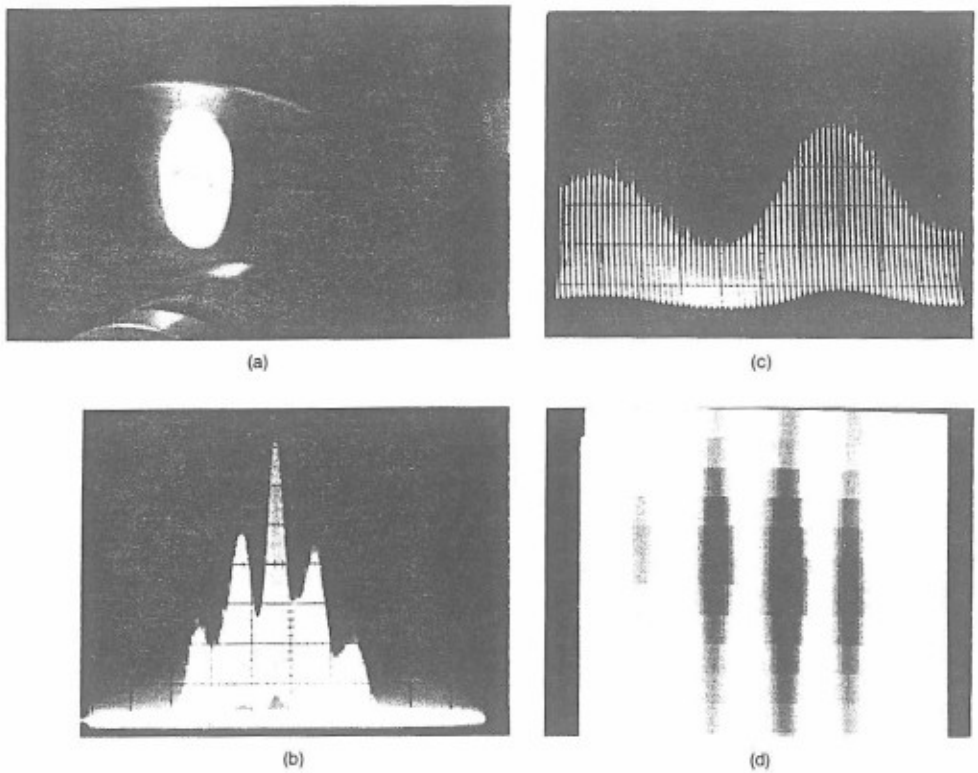


Fig. 5. Experimental demonstration of the system shown in Fig. 4. (a) The pattern recorded on the optical disk. (b) An oscilloscope trace of the detector output. (c) A magnified version of the signal shown in (b) revealing the correlations of the individual lines. (d) The 2-D correlation produced by raster recording the signal shown in (b).

nique has been successfully used in a variety of incoherent architectures.[3]

More than 5 billion bits can be stored in the type of disk that we use for most of our research (a write-once, 12-cm-diameter system from Sony). The number of 100×100 pixel images that can be stored in such a disk is more than 5000, if we assume a generous factor of 100 for loss of space–bandwidth product owing to representation (e.g., area modulation for gray-scale representation). The rate at which all these images can be interrogated for a possible match with the input is limited by one or more of the following factors: the scanning speed of the disk (40 Hz in our case), the speed of the radial scanning mechanism, and the sensitivity and the bandwidth of the output detectors and the electronics following them. As an example consider the system of Fig. 2. At a 40-Hz disk rotation rate we obtain 1000 image correlations per 1/40 of a second (i.e., 40,000 image correlations/sec), which yields a reasonable 4-MHz bandwidth per detector. The input optical power required for reliable detection of the correlation peak is only several milliwatts. It would be extremely difficult to duplicate this capability electronically, and it can be achieved with *existing* optical technology. Moreover it is precisely such capability that is required for practical pattern-recognition problems.

The research reported in this Letter is supported by the U.S. Army Research Office.

References

1. A. Vander Lugt, IEEE Trans. Inf. Theory **IT-10**, 2 (1964).
2. B. D. Guenther, C. R. Christensen, and J. Upatnieks, IEEE J. Quantum Electron. **QE-15**, 1348 (1979).
3. D. Psaltis, Opt. Eng. **23**, 1 (1984).
4. D. Psaltis, E. G. Paek, and S. S. Vankatesh, Opt. Eng. 23, 698 (1984).
5. J. W. Yu, F. H. Mok, and D. Psaltis, Proc. Soc. Photo-Opt. Instrum. Eng. 825, 22 (1987).
6. A. W. Lohmann and H. W. Werlich, Appl. Opt. **10**, 670 (1971).